心理学报 2022, Vol. 54, No. 6, 703−724
Acta Psychologica Sinica

https://doi.org/10.3724/SP.J.1041.2022.00703

认知诊断模型的标准误与置信区间估计：并行自助法*

刘彦楼

(曲阜师范大学教育大数据研究院，山东 济宁 273165)

摘 要 认知诊断模型的标准误(Standard Error, *SE*; 或方差—协方差矩阵)与置信区间(Confidence Interval, CI)在模型参数估计不确定性的度量、项目功能差异检验、项目水平上的模型比较、**Q** 矩阵检验以及探索属性层级关系等领域有重要的理论与实践价值。本研究提出了两种新的 *SE* 和 CI 计算方法：并行参数化自助法和并行非参数化自助法。模拟研究发现：模型完全正确设定时，在高质量及中等质量项目条件下，这两种方法在计算模型参数的 *SE* 和 CI 时均有好的表现；模型参数存在冗余时，在高质量及中等质量项目条件下，对于大部分允许存在的模型参数而言，其 *SE* 和 CI 有好的表现。通过实证数据展示了新方法的价值及计算效率提升效果。

关键词 认知诊断模型，标准误，置信区间，自助法，并行计算

分类号 B841

1 引言

认知诊断模型(Cognitive Diagnosis Model, CDM)或称诊断分类模型，是一类离散潜变量模型(Rupp et al., 2010)，当前已广泛应用于心理、教育或生物学等领域(例如，Tjoe & de la Torre, 2014)。潜在属性在不同领域有不同的含义，例如，知识、技能、认知过程、精神障碍、甚至是病原体等(Rupp et al., 2010; Wu et al., 2017)。恰当应用 CDM，研究者可以通过被试的外显行为去推论每个个体的多维潜在属性掌握状况，为被试提供及时的反馈、个性化的指导或针对性的补救。

CDM 模型参数的标准误(Standard Error, *SE*)是关于模型参数估计不确定性的度量(Liu et al., 2021)。在心理统计与测量模型中，点估计值相同的两个模型参数可能由于 *SE* 不同而具有不同的置信区间(Confidence Interval, CI)，因此需要综合考虑模型参数的点估计值与 CI。例如，CDM 中两个项目的猜测参数估计值均为 0.2，但 *SE* 的估计值分别为 0.08 与 0.05，那么这两个猜测参数的估计精度不同。根据正态分布理论，第一个猜测参数的 95% CI 是[0.2−1.96×0.08, 0.2+1.96×0.08]，第二个猜测参数的 95% CI 是[0.2−1.96×0.05, 0.2+1.96×0.05]。正因如此，国内外多种心理学期刊(如《心理学报》，或参考：American Psychological Association, 2020)要求或建议报告 *SE* 及 95% CI。然而，在国内外的 CDM 实证研究中，报告模型参数的 *SE* 及 CI 的研究仍然较少。造成这种现象的原因是多方面的，主要原因在于缺乏易用的计算方法。接下来，本文将对两类常用的 *SE* 及 CI 的估计方法：解析法以及自助法目前存在的问题展开探讨，并提出一类简易、可行的方法。

CDM 中模型参数的 *SE* (或广义而言，方差—协方差矩阵)在推论统计中具有基础与核心作用(Liu, Xin et al., 2019; Philipp et al., 2018)。除用于计算 CI 外，模型参数的 *SE* 在项目功能差异检验(Liu, Yin, et al., 2019; Ma et al., 2021; 刘彦楼 等, 2016)、项目水平上的模型比较(de la Torre & Lee, 2013;

收稿日期: 2021-10-14
* 国家自然科学基金青年项目(31900794)资助。
通信作者: 刘彦楼, E-mail: liuyanlou@163.com

Liu, Andersson, et al., 2019; Ma & de la Torre, 2016, 2019)、**Q** 矩阵检验(Ma & de la Torre, 2020a)以及探索属性层级关系(Liu et al., 2021; Wang & Lu, 2021)等领域也有重要价值。对于模型参数的 *SE* 的估计，研究者提出了多种基于解析法的估计方法(Liu, Xin et al., 2019; Liu et al., 2021; Philipp et al., 2018; 刘彦楼 等, 2016)，包括：经验交叉相乘信息矩阵法(Empirical Cross-product Information Matrix, XPD)、观察信息矩阵法(Observed Information Matrix, Obs)和三明治信息矩阵法(Sandwich-type Information Matrix, Sw)。

在模型参数可识别条件下(Gu & Xu, 2020; Wang & Lu, 2021)，研究者通过数据模拟以及实证数据分析的方式探索了使用解析法信息矩阵(Liu et al., 2016; 刘彦楼 等, 2016)计算的模型参数(包括项目参数与用于描述被试分布的结构参数)的 *SE* 及 CI 的表现。关于项目参数的 *SE* 及 CI，研究者比较了在理想状况下(即模型与观察数据完美拟合)、在 CDM 的项目反应模型和/或 **Q** 矩阵错误设定条件下，XPD、Obs 或 Sw 方法的表现(Liu, Xin, et al., 2019; Philipp et al., 2018)。研究发现，当模型(包括项目反应模型与 **Q** 矩阵)完全正确设定或存在较少错误设定时，这 3 种方法在项目参数的 *SE* 估计的一致性方面都有好的表现；在模型存在严重错误设定时(如，项目反应模型与 **Q** 矩阵同时包括较多的错误)，只有 Sw 具有健壮性(Liu, Xin, et al., 2019)。关于结构参数的 *SE* 及 CI，研究者在 HCDM (Hierarchical Cognitive Diagnosis Model; Templin & Bradshaw, 2014)框架下进行了探索(Liu et al., 2021)。研究发现，对于正确设定的属性层级关系，即结构模型完全正确设定时，在样本量大于或等于 3000 条件下这 3 种方法均有较好的 95% CI 覆盖率；当属性之间存在层级关系但使用饱和 CDM 估计时，即结构模型参数存在部分冗余情景下，对于允许存在的结构参数(permissible structural parameter)，即根据属性层级关系在理论上不等于 0 的结构参数，XPD 和 Obs 方法计算的 *SE* 有较好的表现；对于非允许存在的结构参数(impermissible structural parameter)，即理论上等于 0 的结构参数，XPD 方法计算的结构参数的 *SE* 表现较好(Liu et al., 2021)。

准确地识别与验证 CDM 中的属性层级关系能够使研究者深入地了解被试作答的心理过程，具有重要的理论与实践价值(Leighton et al., 2004)。然而，实践中预先正确设定属性层级关系是一个非常具有挑战性的过程(Hu & Templin, 2020; Liu et al., 2021; Ma & Xu, 2021; Templin & Bradshaw, 2014; Wang & Lu, 2021)。如果认知诊断测验中存在属性层级关系，使用饱和 CDM 拟合作答反应数据，相应的结构参数近似等于 0。即，饱和 CDM 的结构参数能提供属性层级是否存在的证据(Liu et al., 2021; Templin & Bradshaw, 2014)。Liu 等人(2021)初步提出，结构参数的 *SE* 已知时，可以使用 *z* 统计量探索属性层级关系，具体表达式为，

$$z = \frac{\hat{\eta}}{SE(\hat{\eta})} \tag{1}$$

在公式(1)中 $\hat{\eta}$ 表示结构参数估计值，$SE(\hat{\eta})$ 表示结构参数的 *SE*。

在多数情况下，可以使用 XPD、Obs 或 Sw 方法有效地计算 CDM 中模型参数的 *SE*，但是这些解析性方法主要有两个缺点。(1)需要信息矩阵正定(positive definiteness)。DeCarlo (2011, 2019)发现，CDM 中的边界值问题(boundary problems)，会导致使用信息矩阵计算方差—协方差矩阵时存在非正定问题。关于边界值及其可能导致的信息矩阵非正定问题将在第 2 部分详细阐述。(2)需要方差—协方差矩阵的对角线元素大于 0，如果小于 0 则会导致相应的模型参数的 *SE* 无法计算。然而，在实践中由于计算误差的存在，可能会导致使用信息矩阵求逆计算的方差—协方差矩阵中的某个或某些元素小于 0 (Liu & Maydeu-Olivares, 2014)。例如，第 5 部分实证数据分析中基于 Obs 的方差—协方差矩阵中第 2 个结构参数对应的对角线元素小于 0，而导致 *SE* 无法计算。这也就意味着，如果出现情形(1)，则全部的模型参数的 *SE* 无法计算；如果出现情形(2)，相应的模型参数的 *SE* 无法计算。解析法信息矩阵所存在的以上问题，限制了其理论发展及实践应用。

除解析法外，另一类可用于计算 *SE* 及 CI 的方法是自助法(Davison & Hinkley, 1997; Efron & Tibshirani, 1993)，最常见的有参数化自助法(Parametric Bootstrap, PB)与非参数化自助法(Nonparametric Bootstrap, NPB)。PB 以及 NPB 是一种应用广泛(例如，2019 年 1 月至 2021 年 8 月份发表在《心理学报》上的论文中至少有 20 篇论文用到了自助法)、通用性强，但计算密集(computer-intensive)、费时的方法。与解析法信息矩阵不同，PB 以及 NPB 不需要有较强的前提假设以及大量的公式推导。这类方法是通过 3 个步骤进行的。第一步

是根据观察数据集获得重采样数据集。第二步是根据重采样数据集估计模型参数。以上两步重复进行，直到达到预先设定的重抽样次数。第三步，根据每次重复获得的模型参数估计值，计算 *SE* 以及 CI。PB 与 NPB 的不同之处在于：PB 是先通过观察数据集估计获得模型参数，再使用模型参数模拟生成重采样数据集；NPB 则是通过有放回取样的方式直接从观察数据集中取样。尽管研究者认为自助法可以用于计算 CDM 中的 *SE* 及 CI (Ma & de la Torre, 2020b)，且理论上可以较好地解决解析法信息矩阵在特定条件下无法计算的问题，然而其估计的准确性仍缺乏研究。作为一种计算密集型方法，计算量大、耗时长的缺点不仅限制了 PB 与 NPB 的理论研究，也造成了实践应用的困难。举例而言，在 PB 与 NPB 的应用中，进行重抽样时，如果样本量过少可能会影响到自助法的准确性，如果抽样过多会因计算量增大而影响效率。目前，重抽样次数的选择问题仍存在争议(例如，Bai et al., 2016; Efron & Tibshirani, 1993; Guo & Wind, 2021; Hayes, 2009, 2018; Lai, 2021)。另外，PB 与 NPB 在不同情景中估计 CDM 的模型参数的 *SE* 及 CI 的表现也需要进一步探讨。随着多线程、并行调度等计算技术的发展，并行计算技术被逐步用于计算密集型方法研究(Denwood, 2016; Khorramdel et al., 2019)。仅以自助法为例，Zhang 和 Wang (2020)开发了使用并行自助法的 R 软件包 *bmem*，并探讨了其在统计功效分析中的应用(Zhang, 2014)；线性混合效应模型软件包 *lme4* (Bates et al., 2015)也提供了并行计算的自助法，Jiang 等人(2021)以此为基础探索了使用自助法计算概化系数的 CI 估计值问题。

本文要解决的主要问题有：(1)借鉴以往研究中的并行自助法计算技术，开发适用于 CDM 的并行参数化自助法(parallel Parametric Bootstrap, pPB)和并行非参数化自助法(parallel Nonparametric Bootstrap, pNPB)，提高 CDM 中 PB 与 NPB 的计算效率。(2)系统探讨 pPB 与 pNPB 在估计 CDM 模型参数的 *SE* 及 CI 时的表现。正如本文将要呈现的一样，pPB 与 pNPB 是一类简易、可行的方法，不仅能有效解决 CDM 中 *SE* 与 CI 理论研究中的重要问题，而且能有效提升实践应用中的计算效率。

接下来，本文将首先说明解析法信息矩阵计算 *SE* 时存在的问题，然后详细介绍新提出的 pPB 与 pNPB 方法。第 4 部分是模拟研究，分别探讨 CDM 完全正确设定以及存在属性层级关系条件下这两个方法的表现。第 5 部分是实证数据分析，主要用于说明及展示 pPB 与 pNPB 在估计 CDM 模型参数的 *SE* 时的作用与价值。最后是讨论与结论。

2　解析法信息矩阵及其存在的问题

本部分以同一链接(identity link)下的 G-DINA (Generalized Deterministic Input Noisy Output “AND” gate; de la Torre, 2011)为例，分别呈现 3 种解析法信息矩阵并阐述这些方法在计算 CDM 模型参数的 *SE* 及 CI 时可能会遇到的矩阵非正定，以及方差—协方差矩阵对角线元素可能小于 0 等问题。

2.1　饱和的 CDM

假设在一份认知诊断测验中有 N 个被试，J 个项目，K 个属性，且属性及项目均为二级计分，$N \times J$ 维项目反应矩阵记为 $\mathbf{x} \in \{x_{nj}\}$，$J \times K$ 维 $\mathbf{Q}$ 矩阵记为 $\mathbf{Q} = \{q_{jk}\}$。在饱和的 G-DINA 模型中，被试 n 正确作答项目 j 的概率为，

$$P(x_{nj}=1|\boldsymbol{\alpha}_n,\mathbf{q}_j)=\lambda_{j,0}+\sum_{k=1}^{K}\lambda_{j,1,(k)}\alpha_{nk}q_{jk}+\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K}\lambda_{j,2,(k,k')}\alpha_{nk}\alpha_{nk'}q_{jk}q_{jk'}+\cdots \tag{2}$$

其中，$\boldsymbol{\alpha}_n=(\alpha_{n1},\cdots,\alpha_{nK})'$ 是第 n 个被试的属性掌握模式，$\mathbf{q}_j=(q_{j1},\ldots,q_{jK})'$ 是 $\mathbf{Q}$ 矩阵中所定义的正确作答项目 j 所需要的属性，$\boldsymbol{\lambda}_j=(\lambda_{j,0},\lambda_{j,1,(k)},\cdots)'$ 是项目 j 的所有参数。对于饱和 G-DINA 模型进行恰当约束，可以获得多种特殊模型。

为便于理解及行文，以 $K=2$，$\mathbf{q}_j=(1,1)'$，$\boldsymbol{\alpha}_n=(1,1)'$ 为例。饱和 G-DINA 的项目反应函数可以表达为，

$$P(x_{nj}=1|\boldsymbol{\alpha}_n,\mathbf{q}_j)=\lambda_{j,0}+\lambda_{j,1,(1)}+\lambda_{j,1,(2)}+\lambda_{j,2,(1,2)} \tag{3}$$

其中，$\lambda_{j,0}$ 为截距参数，表示没有掌握项目所需的任何属性仅凭猜测正确作答项目 j 的概率，$\lambda_{j,1,(1)}$ 和 $\lambda_{j,1,(2)}$ 分别是对应于第一个属性(α_1)和第二个属性(α_2)的主效应参数，$\lambda_{j,2,(1,2)}$ 是这两个属性的交互效应。

当 $K=2$ 且属性层级关系不存在时，所有可能的属性掌握模式可以表示为，

$$\boldsymbol{\alpha}=\begin{pmatrix}\boldsymbol{\alpha}_1'\\ \boldsymbol{\alpha}_2'\\ \boldsymbol{\alpha}_3'\\ \boldsymbol{\alpha}_4'\end{pmatrix}=\begin{pmatrix}0&0\\1&0\\0&1\\1&1\end{pmatrix}$$

即 $L=2^K=4$。使用同一链接函数，可以将以上用于描述属性掌握模式分布的结构参数 $\boldsymbol{\eta}=(\eta_1,\cdots,\eta_L)'$ 表示为 $\eta_l=p(\boldsymbol{\alpha}_l)$。因为所有的属性掌握模式概率

之和等于 1，所以将最后一个结构参数约束为 $\eta_L = 1 - \sum_{l=1}^{L-1} \eta_l$。

2.2 带有属性层级关系的 CDM

当测验所测属性之间存在层级关系时，对饱和模型(如 G-DINA)的结构参数以及项目参数加以适当约束，可获得 HCDM (Templin & Bradshaw, 2014)。同样以 $K = 2$，$\mathbf{q}_j = (1,1)'$，$\boldsymbol{\alpha}_n = (1,1)'$ 为例，且假定这两个属性之间存在线性层级关系：只有掌握 α_1 才能掌握 α_2。那么，所有可能的属性掌握模式为，

$$\boldsymbol{\alpha}^* = \begin{pmatrix} \boldsymbol{\alpha}'_1 \\ \boldsymbol{\alpha}'_2 \\ \boldsymbol{\alpha}'_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$$

由于属性层级关系约束，饱和结构模型中的第三种属性掌握模式 $\boldsymbol{\alpha}_3$ 不存在，即 $\eta_3 = p(\boldsymbol{\alpha}_3) = 0$。在当前的例子中，HCDM 的项目反应函数可以表示为，

$$P(x_{nj} = 1 \mid \boldsymbol{\alpha}_n, \mathbf{q}_j) = \lambda_{j,0} + \lambda_{j,1,(1)} + \lambda_{j,2,(1,2)} \quad (4)$$

可以发现，如果真模型是 HCDM，但使用饱和 G-DINA 模型估计参数时，某些结构参数(例如，η_3)以及项目参数(例如，饱和 G-DINA 中的 $\lambda_{j,1,(2)}$)的真值都等于 0，在这种情况下会导致 CDM 中的一些模型参数冗余。在接下来的部分中，参考以往研究中的表述(Liu, 2018; Liu et al., 2021)，将真值为 0 的参数统称为非允许存在的参数，真值不等于 0 的参数统称为允许存在的参数。

2.3 解析法信息矩阵及其不足

在一定的正则性假设下(Bishop et al., 2007)，CDM 模型参数的极大似然估计值向量 $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\lambda}}', \hat{\boldsymbol{\eta}}')'$ 与真值向量 $\boldsymbol{\gamma}$ 的差，服从均值为 $\mathbf{0}$ 向量、方差—协方差矩阵为 $\mathcal{I}_E^{-1}$ 的多元正态分布(Liu et al., 2016)，

$$\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}_E^{-1}) \quad (5)$$

公式(5)中，$\mathcal{I}_E$ 表示的是使用模型参数真值以及对单个被试的作答反应向量求期望(即所有可能的作答反应模式)而计算的期望 Fisher 信息矩阵(Liu et al., 2016; Liu, Xin et al., 2019)。但由于模型参数真值在实践中是未知的，并且所有可能的作答反应模式会随着项目的数量呈现指数增长，因此 $\mathcal{I}_E$ 只具有理论价值，无法应用于实践(Liu, Xin et al., 2019)。

针对 $\mathcal{I}_E$ 的不足，研究者提出使用模型参数估计值 $\hat{\boldsymbol{\gamma}}$ 替代真值 $\boldsymbol{\gamma}$，使用被试的观察作答反应矩阵 $\mathbf{x}$ 代替单个被试的作答反应向量的期望，进而开发出 XPD、Obs 以及 Sw 矩阵(Liu, Xin et al., 2019; Philipp et al., 2018; 刘彦楼 等, 2016)。使用观察数据对数似然函数 $\ell(\hat{\boldsymbol{\gamma}} \mid \mathbf{x})$ 关于模型参数 $\boldsymbol{\gamma} = (\boldsymbol{\lambda}', \boldsymbol{\eta}')'$ 的一阶导向量交叉相乘而计算的 XPD 矩阵的表达式为，

$$\mathcal{I}_{XPD} = \begin{bmatrix} \frac{\partial \ell(\hat{\boldsymbol{\gamma}} \mid \mathbf{x})}{\partial \lambda_1} \frac{\partial \ell(\hat{\boldsymbol{\gamma}} \mid \mathbf{x})}{\partial \lambda_1} & \cdots & \frac{\partial \ell(\hat{\boldsymbol{\gamma}} \mid \mathbf{x})}{\partial \lambda_1} \frac{\partial \ell(\hat{\boldsymbol{\gamma}} \mid \mathbf{x})}{\partial \eta_{L-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \ell(\hat{\boldsymbol{\gamma}} \mid \mathbf{x})}{\partial \eta_{L-1}} \frac{\partial \ell(\hat{\boldsymbol{\gamma}} \mid \mathbf{x})}{\partial \lambda_1} & \cdots & \frac{\partial \ell(\hat{\boldsymbol{\gamma}} \mid \mathbf{x})}{\partial \eta_{L-1}} \frac{\partial \ell(\hat{\boldsymbol{\gamma}} \mid \mathbf{x})}{\partial \eta_{L-1}} \end{bmatrix} \quad (6)$$

根据观察数据对数似然函数关于模型参数的二阶偏导而计算的 Obs 矩阵可表示为(Liu, Xin et al., 2019; 刘彦楼 等, 2016)，

$$\mathcal{I}_{Obs} = -\begin{bmatrix} \frac{\partial^2 \ell(\hat{\boldsymbol{\gamma}} \mid \mathbf{x})}{\partial \lambda_1 \partial \lambda_1} & \cdots & \frac{\partial^2 \ell(\hat{\boldsymbol{\gamma}} \mid \mathbf{x})}{\partial \lambda_1 \partial \eta_{L-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell(\hat{\boldsymbol{\gamma}} \mid \mathbf{x})}{\partial \eta_{L-1} \partial \lambda_1} & \cdots & \frac{\partial^2 \ell(\hat{\boldsymbol{\gamma}} \mid \mathbf{x})}{\partial \eta_{L-1} \partial \eta_{L-1}} \end{bmatrix} \quad (7)$$

需要特别说明的是，Obs 矩阵中的元素也可以等价地表达为(Liu & Maydeu-Olivares, 2014; Liu, Xin et al., 2019)，

$$\frac{\partial^2 \ell(\hat{\boldsymbol{\gamma}} \mid \mathbf{x})}{\partial \gamma_1 \partial \gamma_2} = \frac{\partial \ell(\hat{\boldsymbol{\gamma}} \mid \mathbf{x})}{\partial \gamma_1} \frac{\partial \ell(\hat{\boldsymbol{\gamma}} \mid \mathbf{x})}{\partial \gamma_2} - N \sum_{\upsilon=1}^{\upsilon_O} \frac{p_\upsilon}{f(\mathbf{x}_\upsilon)} \frac{\partial^2 f(\mathbf{x}_\upsilon)}{\partial \gamma_1 \partial \gamma_2} \quad (8)$$

在公式(8)中，γ_1 与 γ_2 分别表示任意一个项目参数(λ)或结构参数(η)；υ_O 是作答反应矩阵 $\mathbf{x}$ 中独特反应模式的数量；p_υ 与 $f(\mathbf{x}_\upsilon)$ 分别代表第 υ 个观察到的独特作答模式所占的实际比例以及期望。Sw 矩阵因其形状而得名，表达式为，

$$\mathcal{I}_{Sw} = \mathcal{I}_{Obs}^{-1} \mathcal{I}_{XPD} \mathcal{I}_{Obs}^{-1} \quad (9)$$

可以发现 Sw 矩阵在计算过程中需要 Obs 及 XPD 矩阵的参与。

基于以上陈述，接下来将重点阐述解析法信息矩阵的不足。首先，边界值问题会对解析法信息矩阵造成严重影响。在 CDM 中，至少有两种情形会导致边界值问题，使得无法使用解析法信息矩阵计算 *SE* 或者使 *SE* 变大(DeCarlo, 2011, 2019)。一种可能的情况是：由于项目参数 $\lambda_{j,0}$ 表示的是截距项参数，其取值范围介于[0, 1]之间。然而，在 $\lambda_{j,0}$ 的真值等于 0 或 1 的极端情况下，由于真值在参数空间的边界上，$\lambda_{j,0}$ 的估计值有较大可能会非常接近 0 或 1，造成项目参数的边界值问题。另一种可能的情况是：CDM 中有非允许存在的结构参数。当 CDM 中存在属性层级关系但使用饱和模型估计的时候，不可避免的有非允许存在的项目参数及结构参数。

因为结构参数的取值区间为[0, 1], 非允许存在的结构参数的真值恰好落在参数空间边界上, 其估计值可能会非常接近 0, 例如, 10^{-6}。边界值问题会造成解析法信息矩阵不稳定或者是奇异阵(Liu et al., 2021)。其次, 如果非允许存在的结构参数的估计值偏离其真值 0, 那么这个估计值是有偏的, 不再符合公式(5)中的前提假设, 因此对 XPD、Obs 以及 Sw 矩阵的计算会造成不良影响。第三, 可以发现, Obs 矩阵等于 XPD 矩阵减去公式(8)中最右侧部分的表达式。但是由于计算误差的存在, Obs 矩阵中对角线元素可能会小于 0, 对应模型参数的 *SE* 无法计算, 这是 Obs 矩阵的一个不足(Liu & Maydeu-Olivares, 2014)。

3　并行非参数化及参数化自助法

3.1　并行非参数化自助法

NPB 的基本思想是模拟从总体中抽取样本的方式而计算模型参数的 *SE*。假定原始作答反应矩阵 $\mathbf{x}$ 是一个“总体”, 采取有放回取样的方式获得新的“样本” (被称为重抽样样本, 记作 $\mathbf{x}^*$)。根据 $\mathbf{x}^*$ 计算模型的参数估计值向量 $\hat{\boldsymbol{\gamma}}^*$。依次循环 *B* 次, 最终计算 *B* 个 $\hat{\boldsymbol{\gamma}}^*$ 的标准差而获得模型参数估计值 $\hat{\boldsymbol{\gamma}}$ 的 *SE*。然而, NPB 存在运行效率低的问题(Ma & de la Torre, 2020b)。

本研究新提出的 pNPB 的具体实施步骤如下:

步骤(1), 确定重抽样的次数 *B*, 设定拟合模型; 检测 CPU 的核心数量, 据此创建并行运算环境中相应数量的副本程序。

步骤(2), 并行抽样阶段。在运算环境的每个副本程序中独立进行如下操作: (a)从原始作答数据 $\mathbf{x}$ 中采取有放回取样方式获得新的样本 $\mathbf{x}^*$; (b)根据预先设定的 CDM 使用 R 语言中的 *GDINA* (Ma & de la Torre, 2020b) 软件包计算模型参数估计值 $\hat{\boldsymbol{\gamma}}^* = (\hat{\boldsymbol{\lambda}}^*, \hat{\boldsymbol{\eta}}^*)'$。在每个副本程序中重复执行(a)与(b)直到达到预先设定的重抽样次数 *B*。

步骤(3), 根据 *B* 次重复抽样中估计获得的模型参数值 $\hat{\boldsymbol{\gamma}}^*$, 计算模型参数的方差—协方差矩阵。将对角线元素开平方, 可以获得模型参数 $\hat{\boldsymbol{\gamma}}$ 的 *SE*。

3.2　并行参数化自助法

PB 的基本思想是使用模型的参数估计值 $\hat{\boldsymbol{\gamma}}$ 作为“总体参数”, 并使用这些参数模拟生成新 *B* 个重抽样“样本” $\mathbf{x}^*$, 通过这些“样本”估计基于重抽样的模型参数估计值 $\hat{\boldsymbol{\gamma}}^*$。

本研究新提出的 pPB 的实施步骤如下:

步骤(1), 除执行 pNPB 中的步骤(1)外, 还需根据原始作答数据 $\mathbf{x}$ 及预先指定的 CDM 估计模型的项目参数及结构参数 $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\eta}})'$。

步骤(2), 参数化并行抽样阶段。在运算环境的每个副本程序中独立进行如下操作: (a)根据结构参数 $\hat{\boldsymbol{\eta}}$ 模拟生成每个被试的属性掌握模式; (b)根据被试属性掌握模式及项目参数 $\hat{\boldsymbol{\lambda}}$ 模拟生成被试在所有项目上的作答反应矩阵 $\mathbf{x}^*$; (c)根据新的作答反应矩阵 $\mathbf{x}^*$ 及预先设定的模型使用 R 语言中的 *GDINA* (Ma & de la Torre, 2020b)软件包重新估计获得模型的项目及结构参数 $\hat{\boldsymbol{\gamma}}^* = (\hat{\boldsymbol{\lambda}}^*, \hat{\boldsymbol{\eta}}^*)'$。并行运行每个副本程序, 重复执行(a)、(b)与(c)直到达到预先设定的重抽样次数 *B*。pPB 中的步骤(3)与 pNPB 中的步骤(3)相同, 不再赘述。

相对于解析法信息矩阵, pNPB 以及 pPB 的优点在于通用性强, 不需要繁琐的公式推导; 不需要严格的前提假设(如, 模型参数估计值渐近正态等); 不涉及矩阵求逆, 受边界值问题影响较小, 尤其适合 CDM 中有非允许存在结构参数情形下 *SE* 及 CI 的计算; 模型参数的方差—协方差矩阵仅需 *B* 个 $\hat{\boldsymbol{\gamma}}^*$ 向量即可计算, 对角线元素不会出现小于 0 的情况。而且, 与传统的 NPB 以及 PB 相比, 本研究提出的 pNPB 以及 pPB 具有运行速度快, 效率高等优点。这使得本研究可以首次实现在 CDM 中较为充分、系统地探讨使用 pNPB 以及 pPB 计算的 *SE* 及 CI 的表现。

4　模拟研究

4.1　研究目的

CDM 完全正确设定或存在边界值问题时, pNPB 以及 pPB 的表现是本研究重点关注的问题。模拟研究的主要目的有两个: (1)探讨在理想条件下, 即模型完全正确设定时, pNPB 和 pPB 在估计 *SE* 以及 CI 时的表现; 并与解析法 XPD、Obs 和 Sw 的表现进行比较。为使结果具有较好的一般性, 数据生成模型及拟合模型均采用同一链接下的饱和 G-DINA 模型。(2)探讨当属性层级关系存在时, 即当模型的结构参数及项目参数均存在非允许存在的参数时, 这两种方法在估计 *SE* 及 CI 时的表现。需要特别说明的是, 属性间存在层级关系时, XPD、Obs 和 Sw 很容易出现无法求逆的问题(Liu et al., 2021), 因此难以在完全相同的模拟条件下比较自助法与解析法的表现。

检索相关文献(例如, Bai et al., 2016; Efron &

Tibshirani, 1993; Guo & Wind, 2021; Hayes, 2009, 2018; Lai, 2021)发现，研究者对于重抽样次数的设置有较大争议，因此如何找到恰当的重抽样次数也是模拟研究关注的问题。

4.2 研究方法

本研究使用 *GDINA* (Ma & de la Torre, 2020b)软件包估计模型参数，参考 *bmem* (Zhang & Wang, 2020)及 *lme4* (Bates et al., 2015)软件包中开源代码自编 pNPB 以及 pPB 代码，解析法信息矩阵 XPD、Obs 和 Sw 估计代码来自 Liu 等人(2021)，感兴趣的研究者可以联系作者获取。为保证各条件下 CDM 模型参数具有可识别性，尤其是属性层级条件下的模型参数的可识别性(Gu & Xu 2019, 2020)，本研究参考 Ma 和 Xu (2021)的实验设计使用图 1 中呈现的 **Q** 矩阵。另外，为清晰地探讨本研究中各自变量对 pNPB 以及 pPB 的影响，假定数据生成模型中每个条件下的结构参数相等，主效应及交互效应相等，以消除模型参数大小对实验结果的影响。使用云主机运行模拟程序，CPU 型号为英特尔 i9-10980XE，18 核 36 线程，每种实验条件组合重复 $R=500$ 次以获得稳定的模拟结果。

具体而言，数据生成模型有两种：饱和 G-DINA 及存在层级关系($\alpha_1 \rightarrow \alpha_2$，$\alpha_1 \rightarrow \alpha_3$)的 HCDM。数据生成模型为饱和 G-DINA 时，*SE* 估计方法有 5 种：XPD、Obs、Sw、pNPB 以及 pPB；数据生成模型为存在属性层级关系的 HCDM 时，*SE* 估计方法有两种：pNPB 以及 pPB。pNPB 以及 pPB 方法的重抽样次数有 4 个水平：200、500、3000 及 5000 次。样本量有两个水平：1000 及 3000。项目质量有 3 个水平：高质量($P(\mathbf{0})=0.1$，$P(\mathbf{1})=0.9$)、中等质量($P(\mathbf{0})=0.2$，$P(\mathbf{1})=0.8$)、低质量($P(\mathbf{0})=0.3$，$P(\mathbf{1})=0.7$)，其中 $P(\mathbf{0})$ 表示仅凭猜测答对的概率，$P(\mathbf{1})$ 表示掌握项目所需要的全部属性的被试正确作答该项目的概率。所有条件下均使用饱和 G-DINA 模型估计模型参数，也就是当数据生成模型同样为饱和 G-DINA 时，模型参数是完全正确设定的；当数据生成模型为 HCDM 时，模型中存在一些真值为 0 的项目参数与结构参数，此时模型参数是冗余的。

4.3 评价指标

使用偏差(BIAS)以及 95% CI 覆盖率评价 *SE* 估计方法的表现。模型参数估计值的 95% CI 为：

$$95\%\ \mathrm{CI}=\left[\hat{\gamma}\pm z_{\left(\frac{0.05}{2}\right)}SE\left(\hat{\gamma}\right)\right]$$

如果模型参数的 95% CI 在区间 $[0.95\pm 1.96\sqrt{0.95(1-0.95)/R}]$ 内，那么可以认为区间估计是准确的，其中 $SE\left(\hat{\gamma}\right)$ 表示的是使用 XPD、Obs、Sw、pNPB 或 pPB 计算的 *SE*。偏差的计算公式为：

$$\mathrm{BIAS}=\frac{\sum_{r=1}^{R}\left[SE\left(\hat{\gamma}_r\right)-SE\left(\gamma\right)\right]}{R}$$

其中 $SE(\gamma)$ 表示 $R=500$ 次重复中获得的 500 个模型参数估计值向量 $\hat{\boldsymbol{\gamma}}^*$ 的标准差。

4.4 模拟结果

4.4.1 CDM 模型参数完全正确设定条件下的实验结果

图 2 与图 3 分别呈现的是 CDM 完全正确设定时，使用 pNPB 以及 pPB 计算的项目参数 95% CI 覆盖率及 *SE* 的 BIAS。在高质量项目条件下，绝大多数项目参数的 95% CI 都落在图中灰线的理论范围内，BIAS 能很好地接近于 0；并且随着样本量的增加这两项评价指标均在变好。在中等质量项目条件下，$N=1000$ 时尽管有少许项目参数的 95% CI 落在理论范围外且 *SE* 的 BIAS 稍有波动，但绝大部分表现较好，这两个评价指标的波动明显高于高质量项目条件；$N=3000$ 条件下，尤其是 $B\geqslant 500$ 时，绝大多数项目参数的 95% CI 覆盖率以及 *SE* 的 BIAS 控制均有好的表现。在低质量项目条件下，使用 pNPB 以及 pPB 计算的项目参数的 95% CI 覆盖率以及 *SE* 的 BIAS 表现差异明显：在 $N=1000$ 的条件下，使用 pNPB 计算的项目参数的 *SE* 绝大部分在理论区间之上且倾向于高估 *SE*，使用 pPB 计算的项目参数的 *SE* 绝大部分在理论区间之下且会倾向于低估 *SE*；另外可以发现随着样本量的增大，在

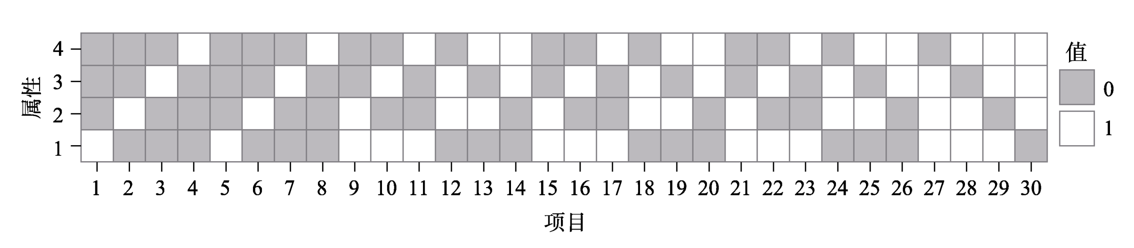


图 1 模拟研究中使用的 **Q** 矩阵

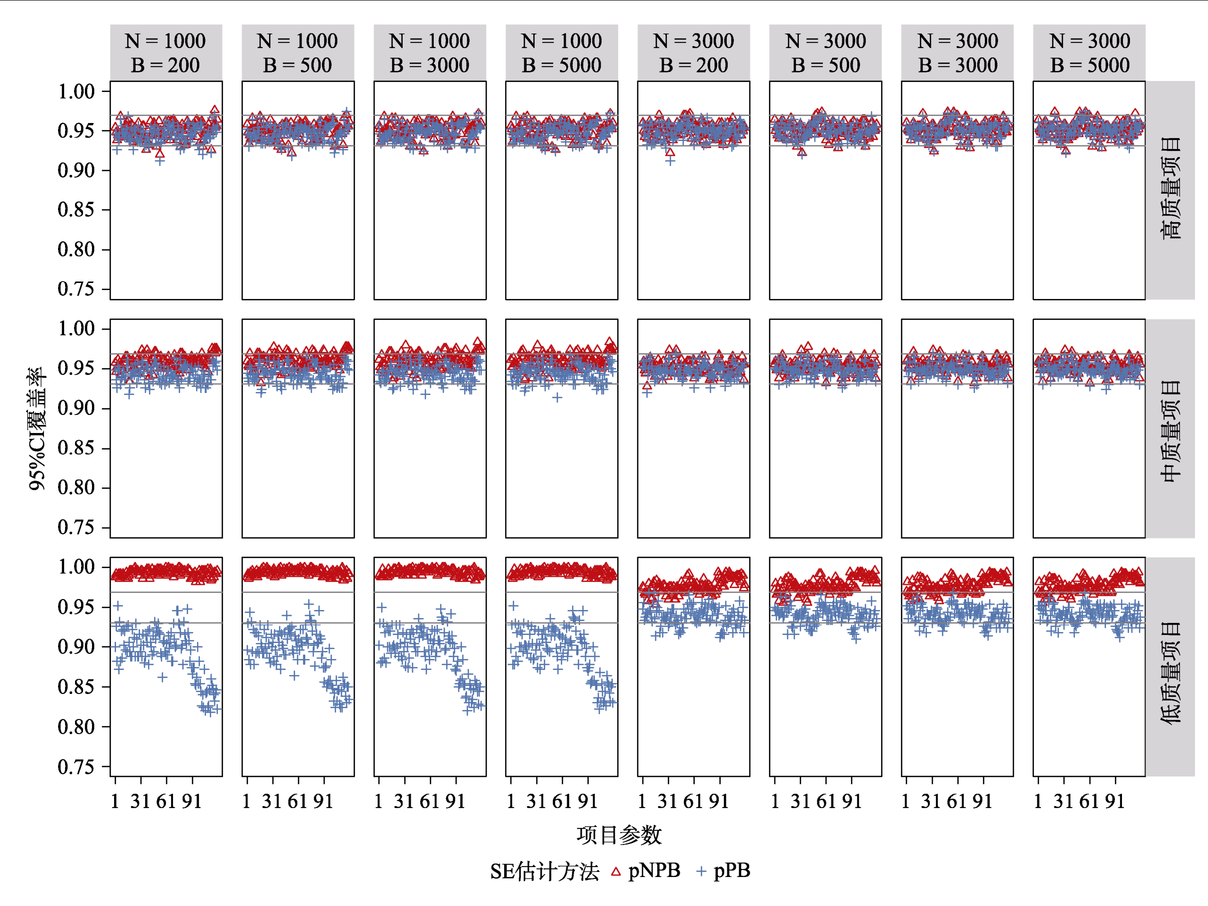


图 2　CDM 模型参数完全正确设定时，基于 pNPB 与 pPB 的项目参数的 95% CI 覆盖率

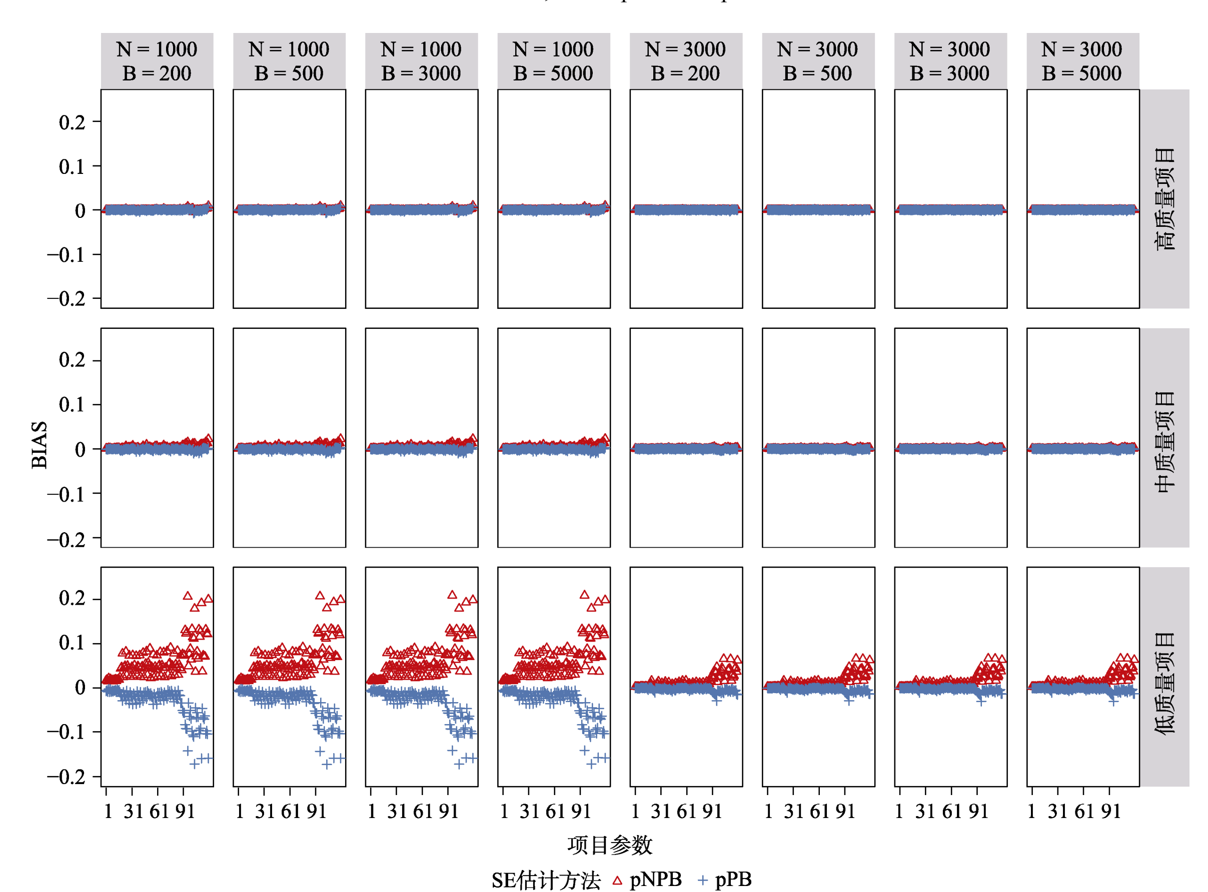


图 3　CDM 模型参数完全正确设定时，基于 pNPB 与 pPB 的项目参数的 *SE* 的 BIAS

$N=3000$条件下项目参数 95% CI 覆盖率及 *SE* 的 BIAS 的表现均在变好，且 pPB 方法的表现优于 pNPB。可以发现，当重抽样次数 $B\geqslant 500$时，相同条件组合下的模拟结果具有高一致性，尤其是 $B=3000$与$B=5000$两者之间没有发现明显差异。

图4与图5呈现的是CDM完全正确设定时，基于解析法 XPD、Obs 与 Sw 的项目参数的 95% CI 覆盖率及 *SE* 的 BIAS。可以发现，高质量以及中等质量项目条件下的项目参数的 *SE* 有好的表现；$N=1000$时，Sw 矩阵的表现略微优于 XPD 与 Obs;

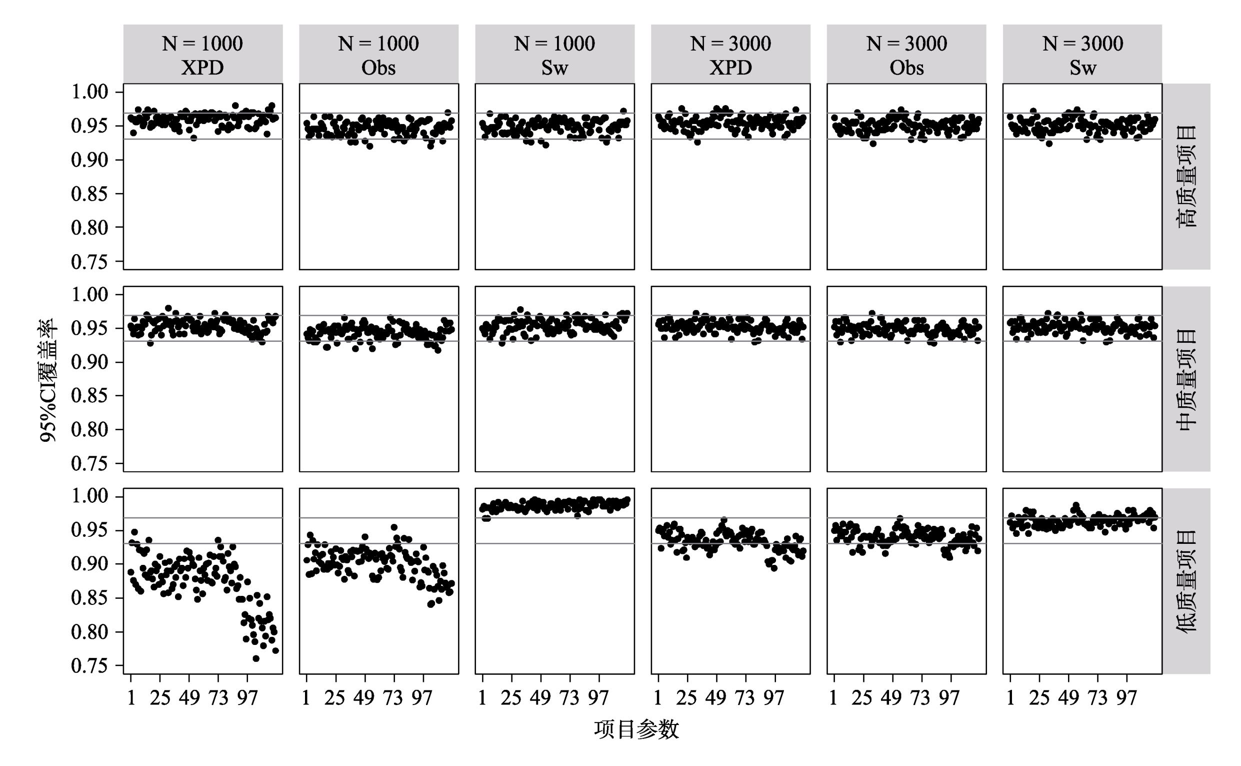


图 4 CDM 模型参数完全正确设定时，基于 XPD、Obs 与 Sw 的项目参数的 95% CI 覆盖率

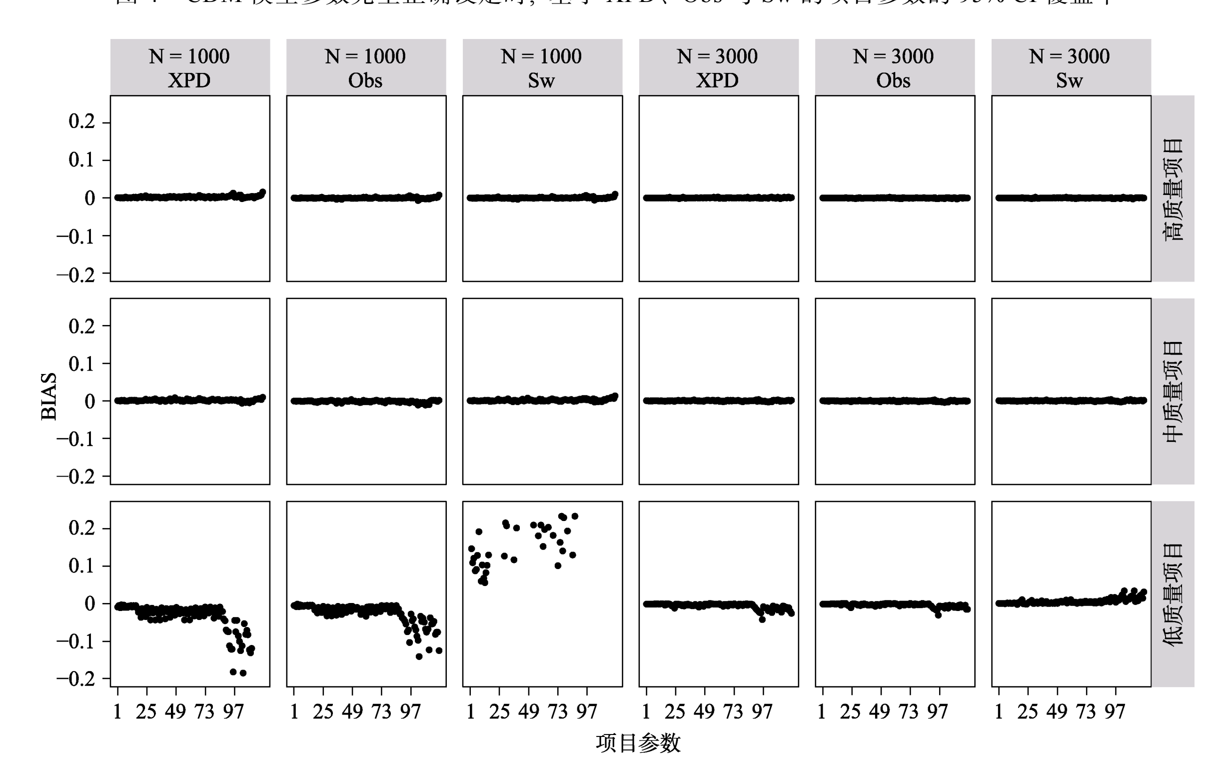


图 5 CDM 模型参数完全正确设定时，基于 XPD、Obs 与 Sw 的项目参数的 *SE* 的 BIAS

当样本量增加到 $N = 3000$ 时，XPD、Obs 以及 Sw 矩阵的表现均在变好。对比高质量以及中等质量项目条件下 XPD、Obs、Sw、pNPB 以及 pPB 的模拟结果，可以发现多数情况下 Sw 以及 Obs 矩阵的表现略微优于其他方法。低质量项目条件下，XPD、Obs 以及 Sw 矩阵计算的项目参数的 *SE* 的表现受到较为严重的影响；$N = 1000$ 时，XPD 与 Obs 的 95% CI 覆盖率绝大部分在理论区间之下且会倾向于低估 *SE*，Sw 的 95% CI 覆盖率绝大部分在理论区间之上且会倾向于高估 *SE*；$N = 3000$ 时，基于 XPD、Obs 以及 Sw 的 95% CI 覆盖率大部分在理论区间内。本研究还发现，低质量项目条件下的 BIAS 结果中，基于 XPD 及 Sw 方法的项目参数的 *SE* 的结果分别有 9 个及 86 个在区间[−0.2, 0.2]之外；检查发现，基于 XPD 及 Sw 方法计算的 *SE* 中有数值极端偏离正常值的结果(例如，*SE* 估计值大于 1000)。这也就是说，在低质量项目且 $N = 1000$ 条件下，XPD 及 Sw 方法的表现不稳定。综合对比低质量项目条件下，XPD、Obs、Sw、pNPB 以及 pPB 的表现，可以发现 Obs 略优于其他方法。

图 6 与图 7 分别呈现的是 CDM 完全正确设定时，基于自助法的结构参数的 95% CI 覆盖率及 *SE* 的 BIAS。可以发现，在高项目质量条件下，使用 pNPB 以及 pPB 计算的结构参数的 *SE* 均有好的表现，所有结构参数的 95% CI 覆盖率都落在图中灰线的理论范围内或边界上，BIAS 几乎完全与 0 重合。在中等质量项目条件下，当 $N = 1000$ 时，尽管结构参数的 95% CI 的波动明显增大，但是大多数结构参数的 *SE* 都有好的表现，且 BIAS 波动也很小；当 $N = 3000$ 时，结构参数的 *SE* 的两种计算方法都有好的表现。在低质量项目条件下，结构参数的 95% CI 覆盖率以及 BIAS 的表现受到严重影响，当 $N = 1000$ 时，绝大多数使用 pNPB 计算的结构参数 95% CI 在理论范围之上且 BIAS 明显大于 0，使用 pPB 计算的 95% CI 全部在理论范围之下且 BIAS 明显小于 0，重抽样次数的增加对于 pNPB 及 pPB 的表现没有明显影响；当 $N = 3000$ 时结构参数的 95% CI 覆盖率及 BIAS 这两个评价指标均在变好，并且可以发现当 $B \geqslant 3000$ 时 pPB 的表现略微优于其他重抽样次数下的表现；但是重抽样次数的增加对于 pNPB 的影响较小。

图 8 与图 9 中呈现的是 CDM 完全正确设定时，基于解析法的结构参数的 95% CI 覆盖率及 *SE* 的 BIAS。在高和中等项目质量条件下，使用 XPD、

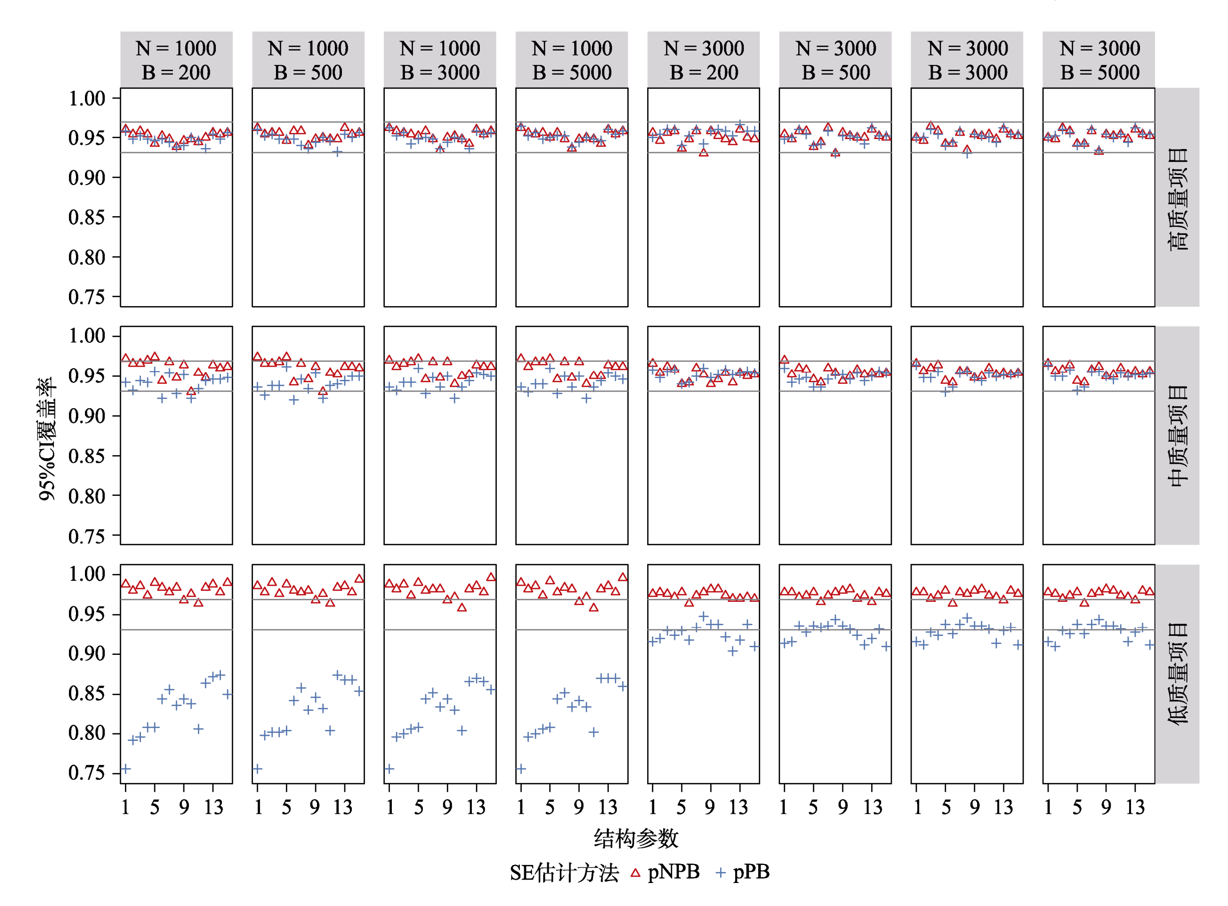


图 6　CDM 模型参数完全正确设定时，基于 pNPB 与 pPB 的结构参数的 95% CI 覆盖率

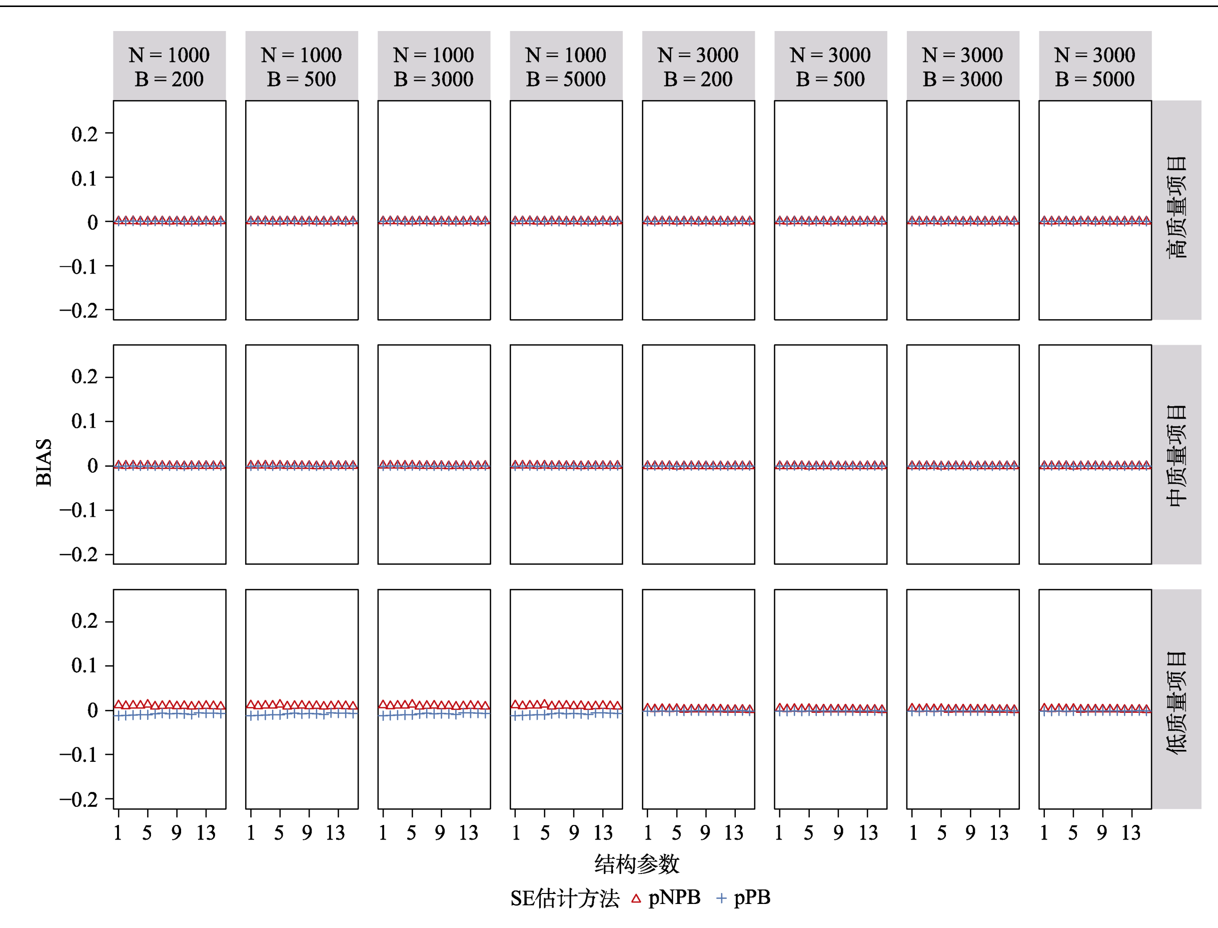


图 7 CDM 模型参数完全正确设定时，基于 pNPB 与 pPB 的结构参数的 *SE* 的 BIAS

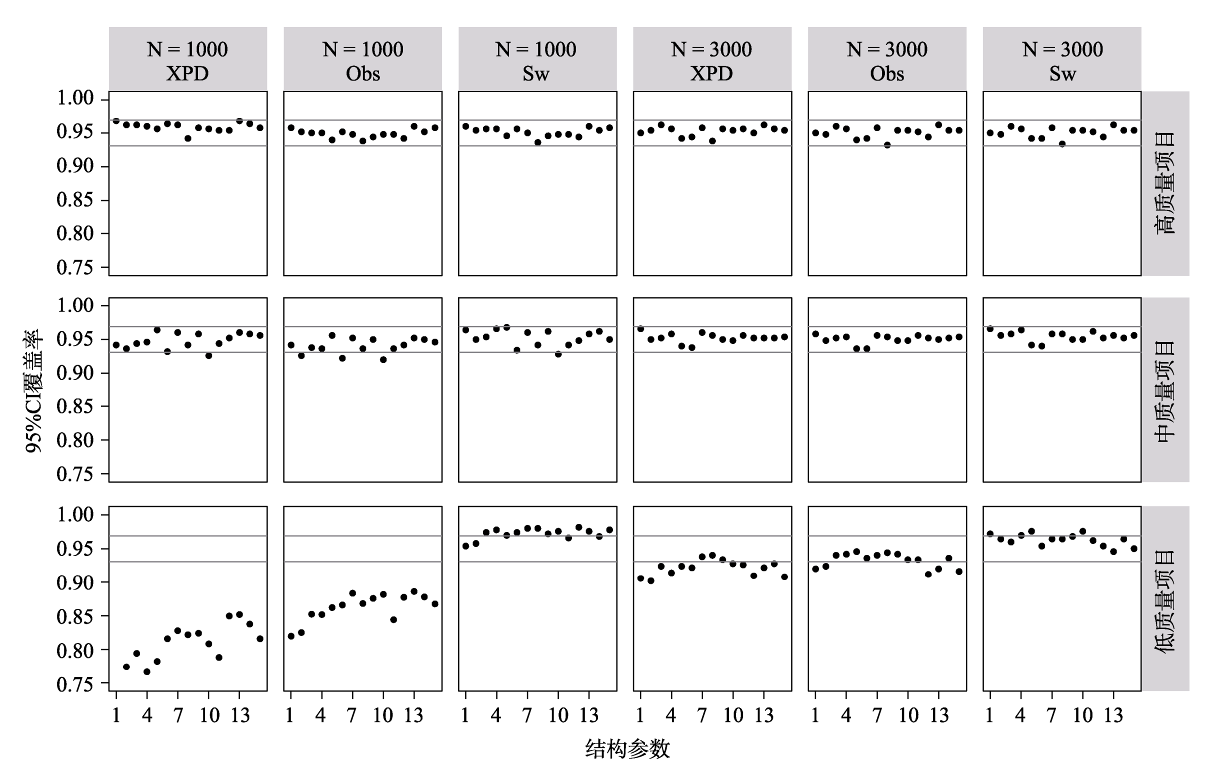


图 8 CDM 模型参数完全正确设定时，基于 XPD、Obs 与 Sw 的结构参数的 95% CI 覆盖率

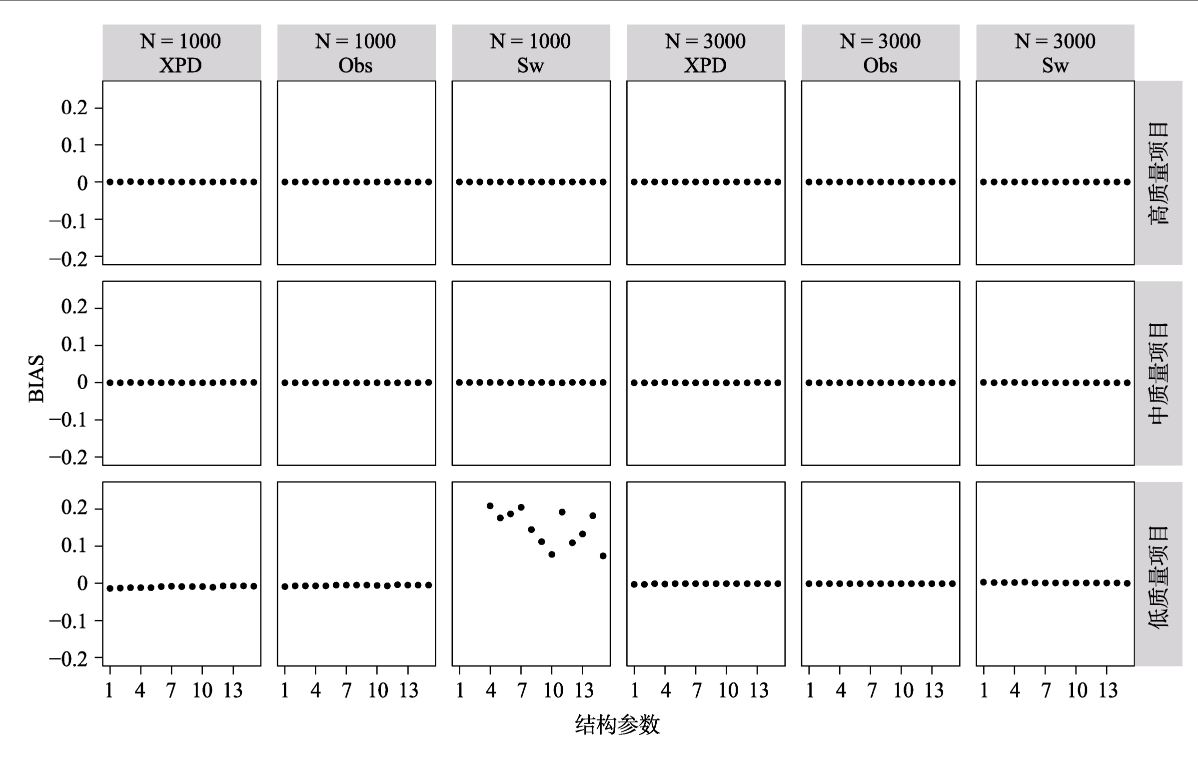


图 9　CDM 模型参数完全正确设定时，基于 XPD、Obs 与 Sw 的结构参数的 *SE* 的 BIAS

Obs 以及 Sw 计算的结构参数的 *SE* 均有好的表现，几乎所有结构参数的 95% CI 覆盖率都落在图中灰线的理论范围内或边界上，BIAS 几乎完全与 0 重合。低质量项目严重影响了使用 XPD、Obs 以及 Sw 计算的结构参数的 *SE* 的表现；$N = 1000$ 时，使用 XPD、Obs 计算的结构参数 95% CI 在理论范围之下且大多数 BIAS 小于 0，使用 Sw 计算的 95% CI 大部分在理论范围之上且 BIAS 明显大于 0；$N = 3000$ 时 XPD、Obs 以及 Sw 计算的结构参数 95% CI 覆盖率及 BIAS 的表现均在变好，尤其是使用 Sw 计算的结构参数 95% CI 大部分在理论范围内。另外，低质量项目且 $N = 1000$ 条件下，基于 Sw 方法计算的结构参数的 95% CI 覆盖率及 BIAS 中分别有 1 个及 3 个值在图 8 及图 9 的区间之外；检查发现，与先前一样，也是由于基于 Sw 方法计算的 *SE* 中有数值极端偏离正常值的结果。综合对比 XPD、Obs、Sw、pNPB 以及 pPB，可以发现除了低质量项目且 $N = 1000$ 条件下以上方法表现均比较差之外，Sw 方法的表现与其他方法相当或优于其他方法。

4.4.2　CDM 的模型参数冗余条件下的实验结果

如前所述，当数据生成模型是 HCDM，但使用饱和模型(如饱和 G-DINA)估计模型参数时，可能会导致模型参数估计值的边界值问题，造成解析法信息矩阵无法求逆或者会产生不稳定的 *SE* 估计结果。自助法不存在矩阵求逆问题，但这种情况下 pNPB 以及 pPB 的表现有待进一步探索。

在模型参数冗余条件下，按照允许存在参数及非允许存在参数这两类分别呈现项目参数及结构参数的 *SE* 的研究结果。另外，为完整显示全部结果，将模型参数冗余条件下的 95% CI 覆盖率的坐标范围设置为[0.3, 1]。图 10 与图 11 呈现的是允许存在项目参数的 95% CI 覆盖率及 *SE* 的 BIAS。可以发现，尽管在高质量及中质量项目条件下，绝大多数的项目参数有良好的 95% CI 覆盖率及 BIAS 控制水平，但是有些参数的 95% CI 低于图中灰线的理论区间，并且存在较大的 BIAS；且在项目质量的所有水平下，这些极端偏离理论区间的项目参数的表现并没有随着其他实验条件的改变而发生明显的变化，甚至在 $N = 3000$ 时更加偏离理论区间。这主要是因为当使用饱和模型估计 HCDM 时，由于错误地设定某些“非允许存在”的属性掌握模式为“存在”，造成了项目参数估计值存在偏差，影响了这些项目参数的 95% CI 覆盖率及 BIAS 表现。例如，对比公式(3)和(4)，可以发现如果“真”模型是带有线性层级关系的 HCDM，但使用饱和 CDM 估计模型参数时，由于“非允许存在”的属性掌握模式 $\boldsymbol{\alpha}_3$ 被错误地设定为“存在”，造成饱和 CDM 中结构参数 η_3 以及项目参数 $\lambda_{j,1,(2)}$ 真值都等于 0。除了极端偏离理论区间的项目参数外，仔细对比高质量及中质量项目条件下理论区间附近的项目参数，可以

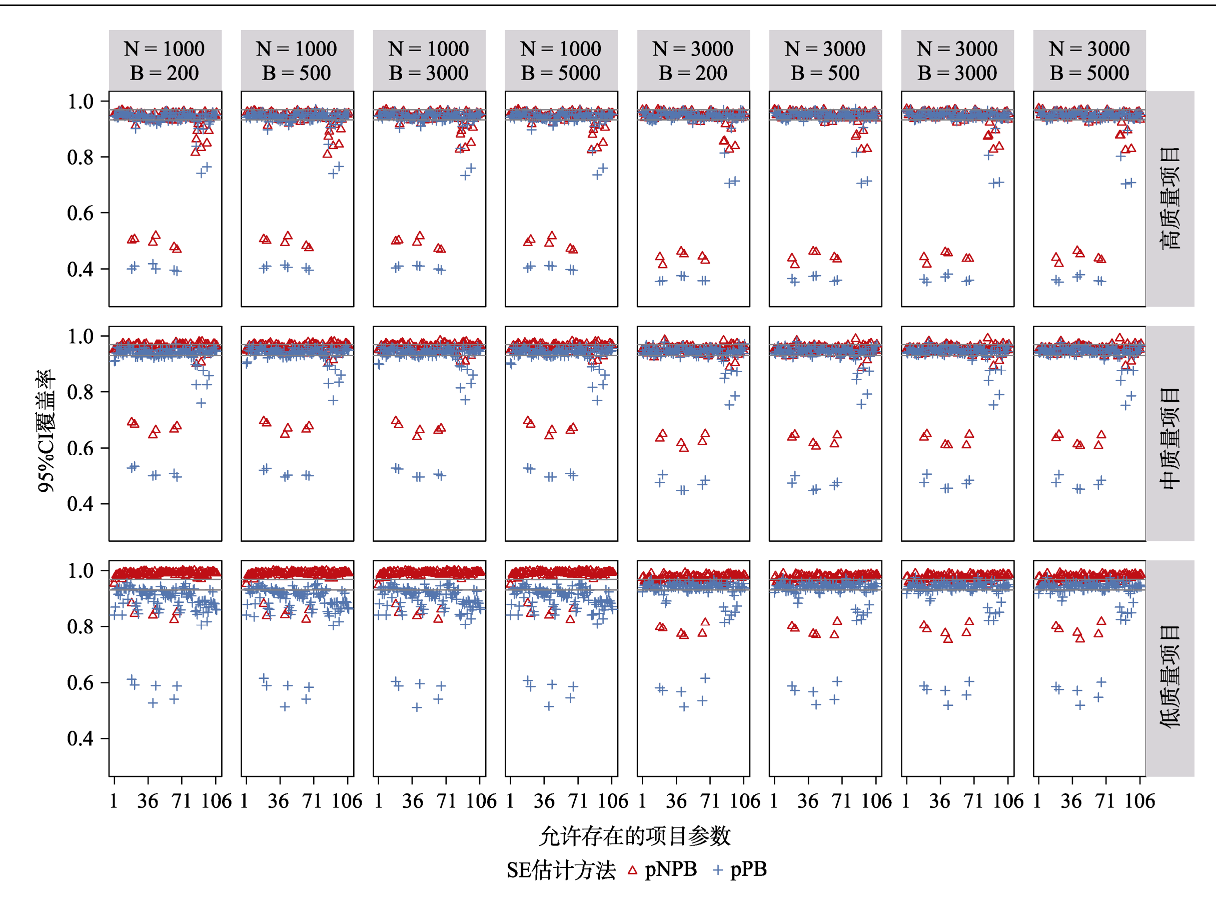


图 10 CDM 模型参数冗余时，基于 pNPB 与 pPB 的允许存在项目参数的 95% CI 覆盖率

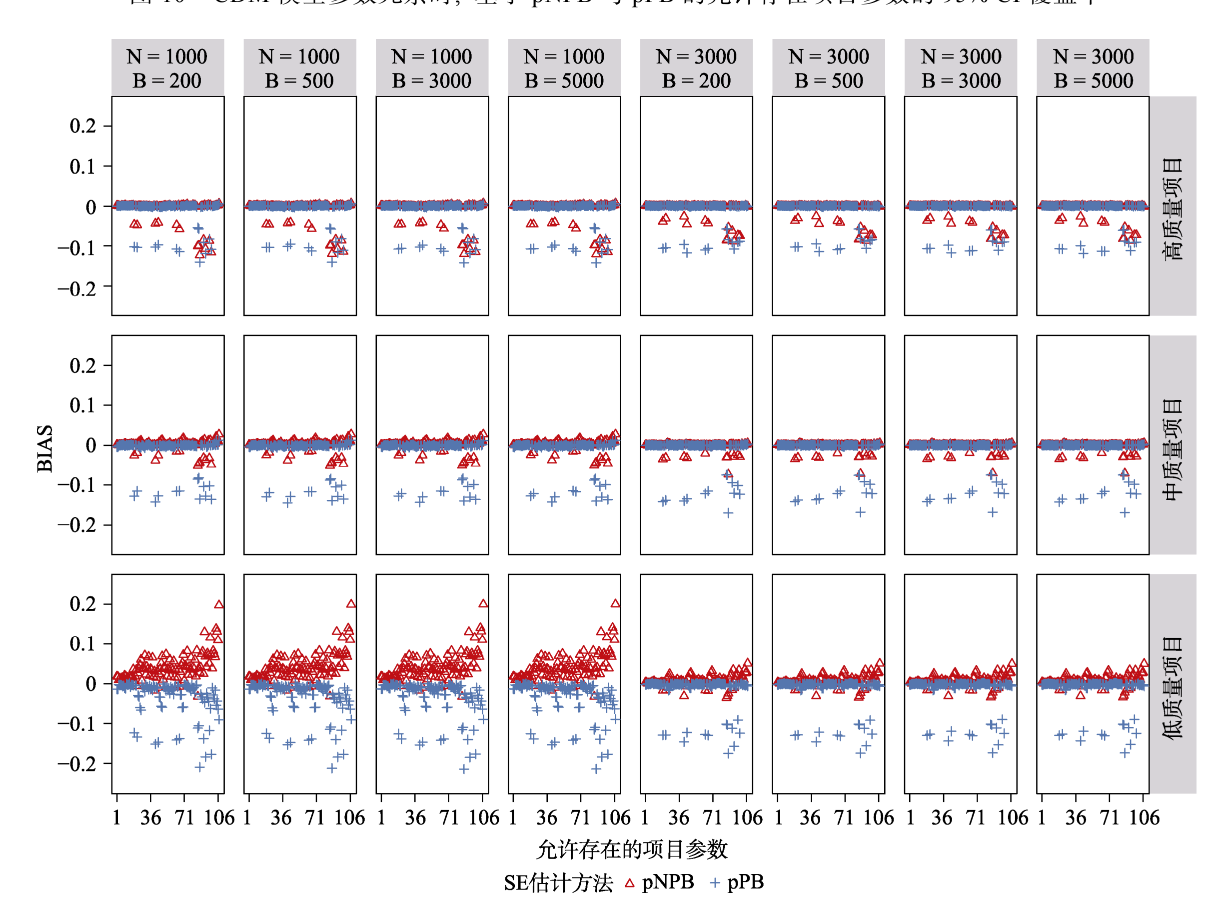


图 11 CDM 模型参数冗余时，基于 pNPB 与 pPB 的允许存在项目参数的 *SE* 的 BIAS

发现随着重抽样次数 B 从 200 增加到 3000, 项目参数的 95% CI 覆盖率略微变好, 但是 $B=3000$ 与 $B=5000$ 两个水平下的结果高度一致。在低质量项目条件下, 允许存在项目参数的 95% CI 覆盖率结果波动明显。

图 12 与图 13 中呈现的是 CDM 模型参数冗余条件下非允许存在项目参数的 95% CI 覆盖率及 *SE* 的 BIAS。整体而言, 大部分非允许存在项目参数的 95% CI 覆盖率低于理论区间, 大部分的 BIAS 值也低于 0。并且在同一个项目质量水平下, 这些非允许存在项目参数的表现具有较高的一致性。另外可以发现样本量、项目质量以及重抽样次数对于这两个指标没有明显影响。从估计方法角度而言, pNPB 在估计非允许存在项目参数的 *SE* 的表现要稍微优于 pPB。

图 14 与图 15 中呈现的是 CDM 模型参数冗余条件下允许存在结构参数的 95% CI 覆盖率和 *SE* 的 BIAS 结果。对于允许存在结构参数而言, 在高质量及中等质量项目条件下, pNPB 及 pPB 方法估计的 95% CI 均在理论区间内或边界线上, 且随着样本量及重抽样次数的增加也在逐渐变好, 允许存在结构参数的 *SE* 的 BIAS 也几乎完全与 0 重合。项目质量对于结构参数的 95% CI 覆盖率及 BIAS 影响明显, 可以发现随着项目质量降低结构参数 95% CI 覆盖率的波动明显增大, BIAS 对于 0 的偏离也在增大。在低质量项目条件下, 当 $N=1000$ 时使用 pPB 估计的结构参数的 95% CI 覆盖率全部在理论区间之下, 且通过 BIAS 结果可以发现此种情况下 pPB 倾向于低估 *SE*; 使用 pNPB 估计的结构参数 95% CI 覆盖率多数在理论区间之上, 且通过 BIAS 结果可以发现这种方法倾向于高估 *SE*; 另外可以发现增加样本量可以改进 pNPB 和 pPB 的表现, 但是增加重抽样次数几乎没有影响。

图 16 与图 17 中呈现的是非允许存在结构参数的 95% CI 覆盖率和 *SE* 的 BIAS 结果。正如本文先前所述, 冗余结构参数的存在对项目参数估计值产生了影响, 进而影响到项目参数的 *SE* 的表现。因此, 如何消除非允许存在结构参数是一个非常有价值的问题。先前研究(Liu et al., 2021)探讨了使用解析法计算 *SE*, 然后通过公式(1)中呈现的 z 统计量对结构参数进行显著性检验的方法消除非允许存在结构参数。通过 z 统计量公式可以发现, 准确的结构参数的 *SE*, 即 $SE(\hat{\eta})$, 对消除非允许存在结构参数特别重要。但是解析法存在边界值及奇异矩阵问题, 影响了 XPD、Obs 及 Sw 的实践应用。pNPB 以及 pPB 不存在以上不足, 因此使用这两种方法计算的非允许存在结构参数的 *SE* 的表现需要重点关注。通过图 16 可以发现, 非允许存在结构参数的 95% CI 覆盖率受到项目质量的影响。在高质量项目条件下使用 pNPB 以及 pPB 计算的 95% CI 覆盖率均稍微高于理论区间。出现这种情况的原因主要在于, 在高质量项目条件下所获的 B 个 $\hat{\boldsymbol{\gamma}}^*$ 的标准差 $SE(\hat{\eta})$ 大于在 R 次重复中获得的 $\hat{\boldsymbol{\gamma}}$ 的标准差 $SE(\eta)$, 即高质量项目条件下结构参数受重抽样因素的影响而产生的变化相对更大。但是通过与图 17 中的 BIAS 结果进行对照可以发现, 整体而言, 通过 pNPB 以及 pPB 估计的 *SE* 与通过多次重复中的模型参数而计算的标准差的绝对差异非常小; 另外 pNPB 以及 pPB 估计的 *SE* 表现在其他条件下的差异很小, 尤其是增加重抽样次数没有改善这两种方法的表现。

在中等质量项目条件下, 基于 pNPB 的非允许存在结构参数的 95% CI 覆盖率大部分在理论区间之上, 基于 pPB 的 95% CI 覆盖率大部分在理论区间之内。即, 中等质量项目条件下通过非参数方法获得重抽样样本 $\mathbf{x}^*$ 而计算的 $\hat{\boldsymbol{\gamma}}^*$ 的 $SE(\hat{\eta})$ 与 $SE(\eta)$ 更为接近, 因此 pNPB 的表现相对较好。随着样本量的增大, 除了使用 pPB 计算的第三个结构参数的 *SE* 外, 其余均更接近理论区间。可以发现, 增加重抽样次数同样没有改善这两种方法的表现。

在低质量项目条件下, 样本量大小对于非允许存在结构参数的 *SE* 表现的影响明显。当 $N=1000$ 时, 基于 pNPB 的 95% CI 覆盖率高于理论区间, 基于 pPB 的 95% CI 覆盖率则几乎全部都低于理论区间。出现以上表现的原因主要是, 相对于 $SE(\eta)$ 而言, 非参数方法获得重抽样样本 $\mathbf{x}^*$ 而计算的 B 个 $\hat{\boldsymbol{\gamma}}^*$ 的 $SE(\hat{\eta})$, 在多数情况下相对更大。但是随着样本量的增加($N=3000$)这两种方法在 95% CI 覆盖率及 BIAS 上的表现也在变好。另外, 将重抽样次数从 $B=200$ 增加到 $B=5000$ 对 *SE* 的表现几乎没有任何影响。

5　实证数据分析

在 CDM 研究中, ECPE (the Examination for the Certificate of Proficiency in English; Templin & Bradshaw, 2014)是经典的实证数据之一。本研究所用 ECPE 数据通过 *CDM* (Robitzsch et al., 2020)软件包公开获取, 包含 2922 名被试在 28 个二值计分的

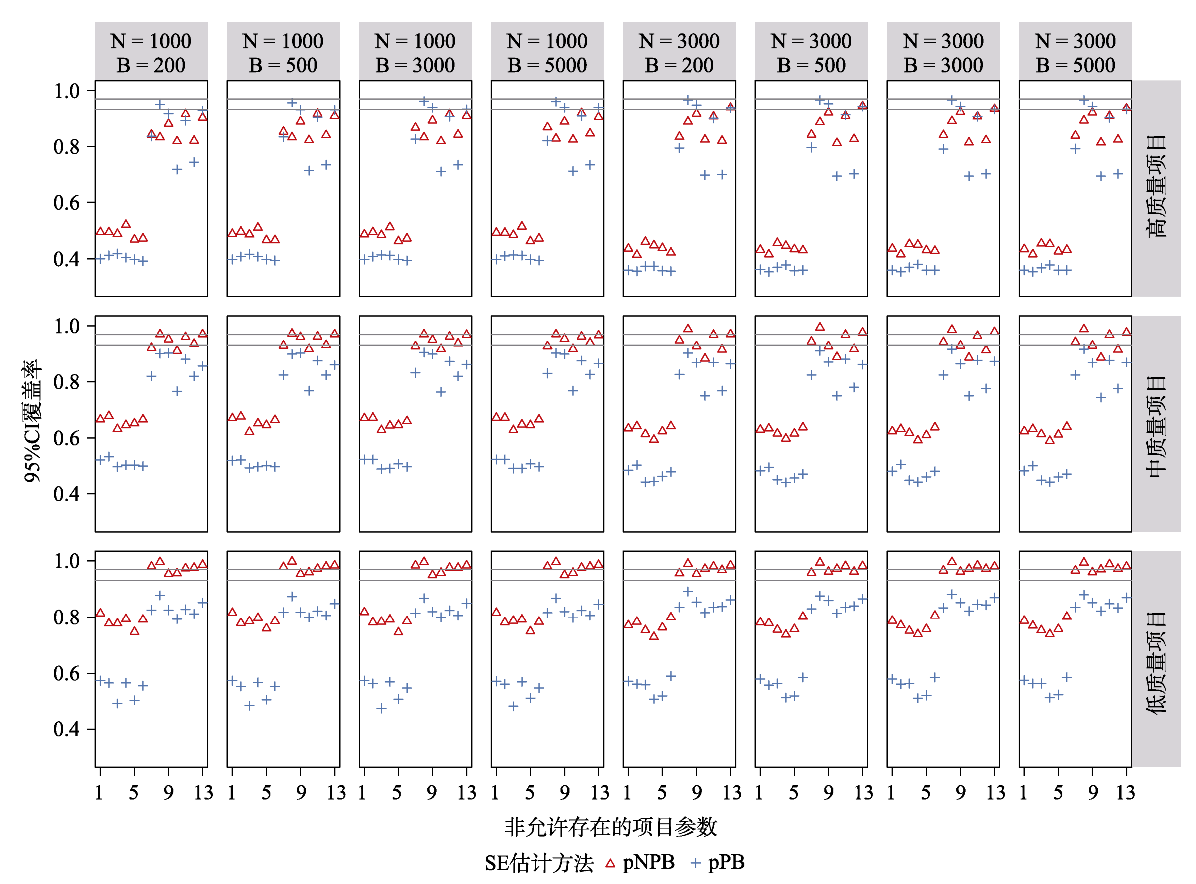


图 12 CDM 模型参数冗余时，基于 pNPB 与 pPB 的非允许存在项目参数的 95% CI 覆盖率

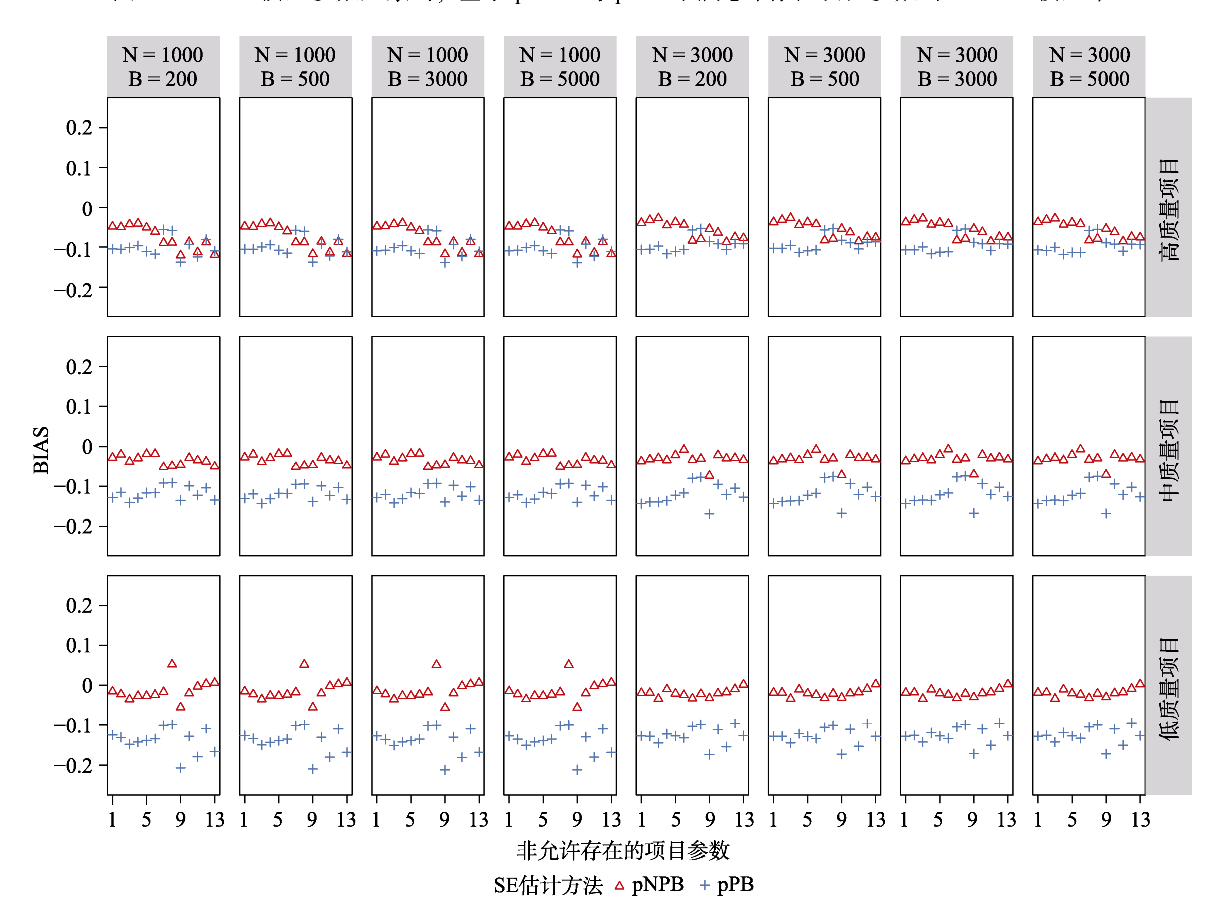


图 13 CDM 模型参数冗余时，基于 pNPB 与 pPB 的非允许存在项目参数的 *SE* 的 BIAS

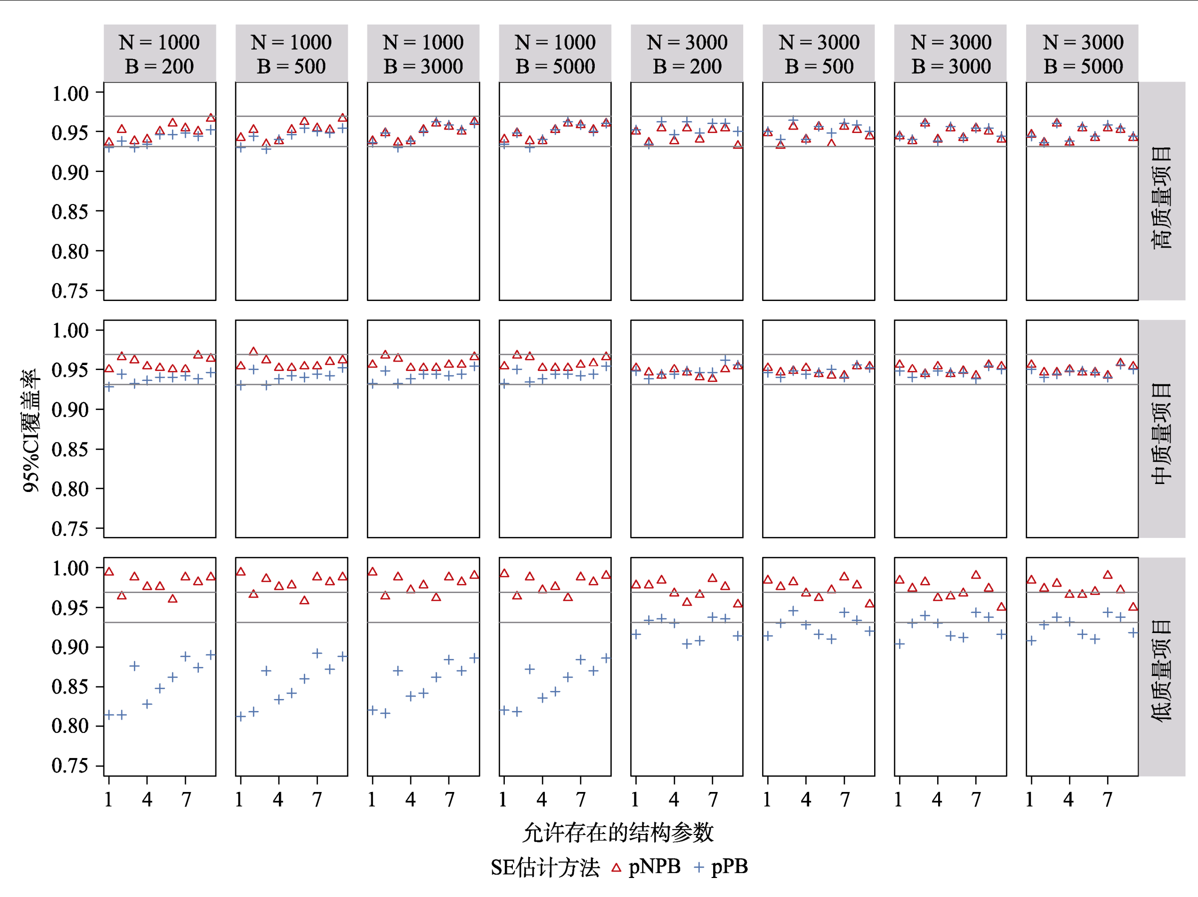


图 14　CDM 模型参数冗余时，基于 pNPB 与 pPB 的允许存在结构参数的 95% CI 覆盖率

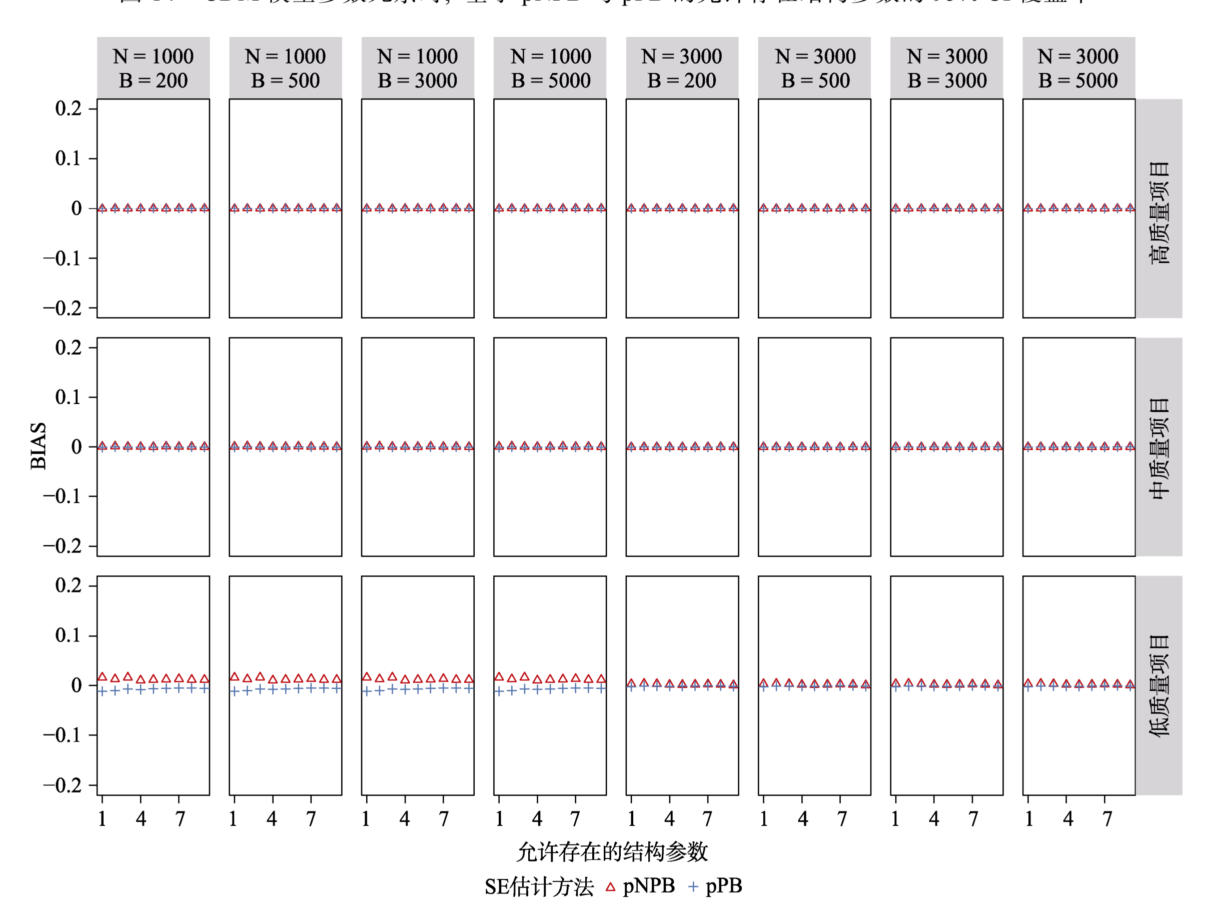


图 15　CDM 模型参数冗余时，基于 pNPB 与 pPB 的允许存在结构参数的 *SE* 的 BIAS

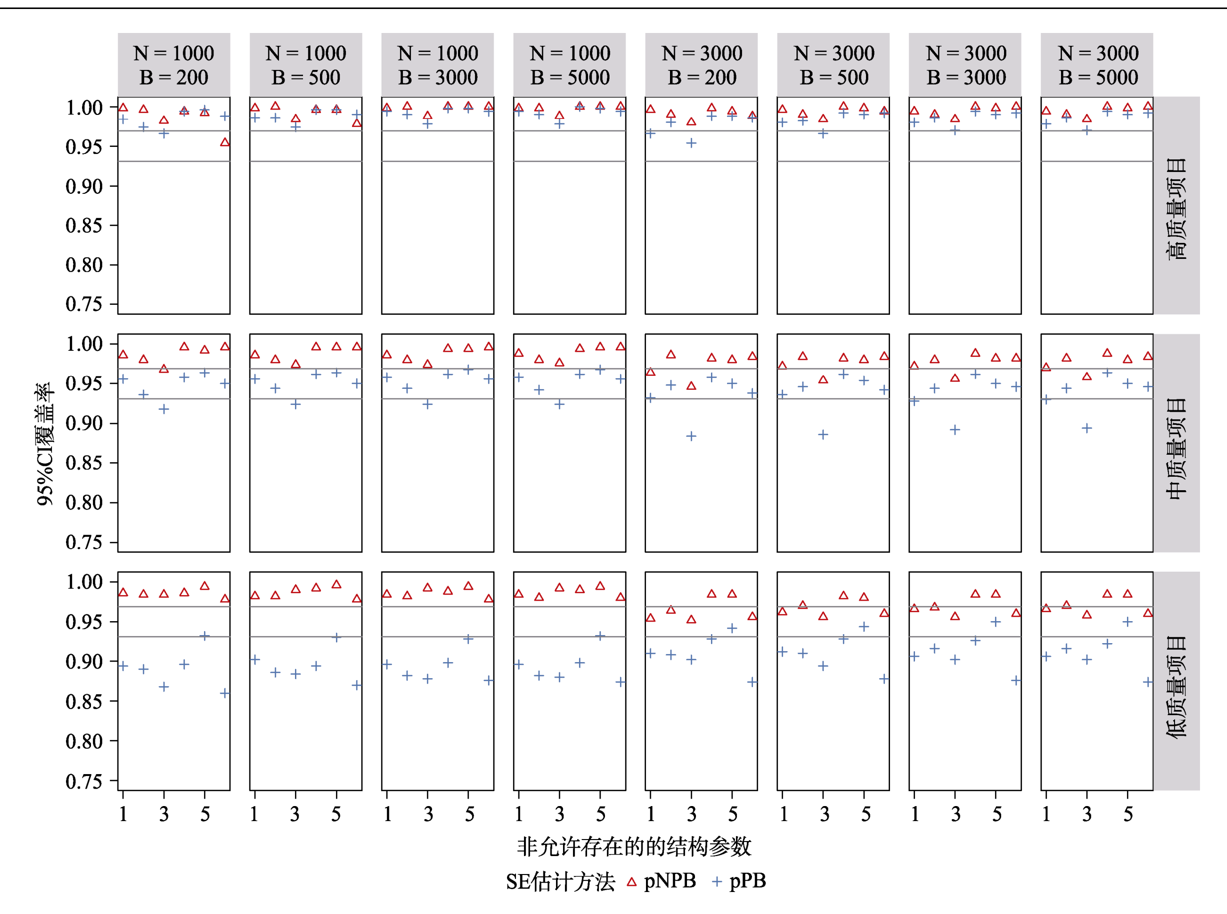


图 16　CDM 模型参数冗余时，基于 pNPB 与 pPB 的非允许存在结构参数的 95% CI 覆盖率

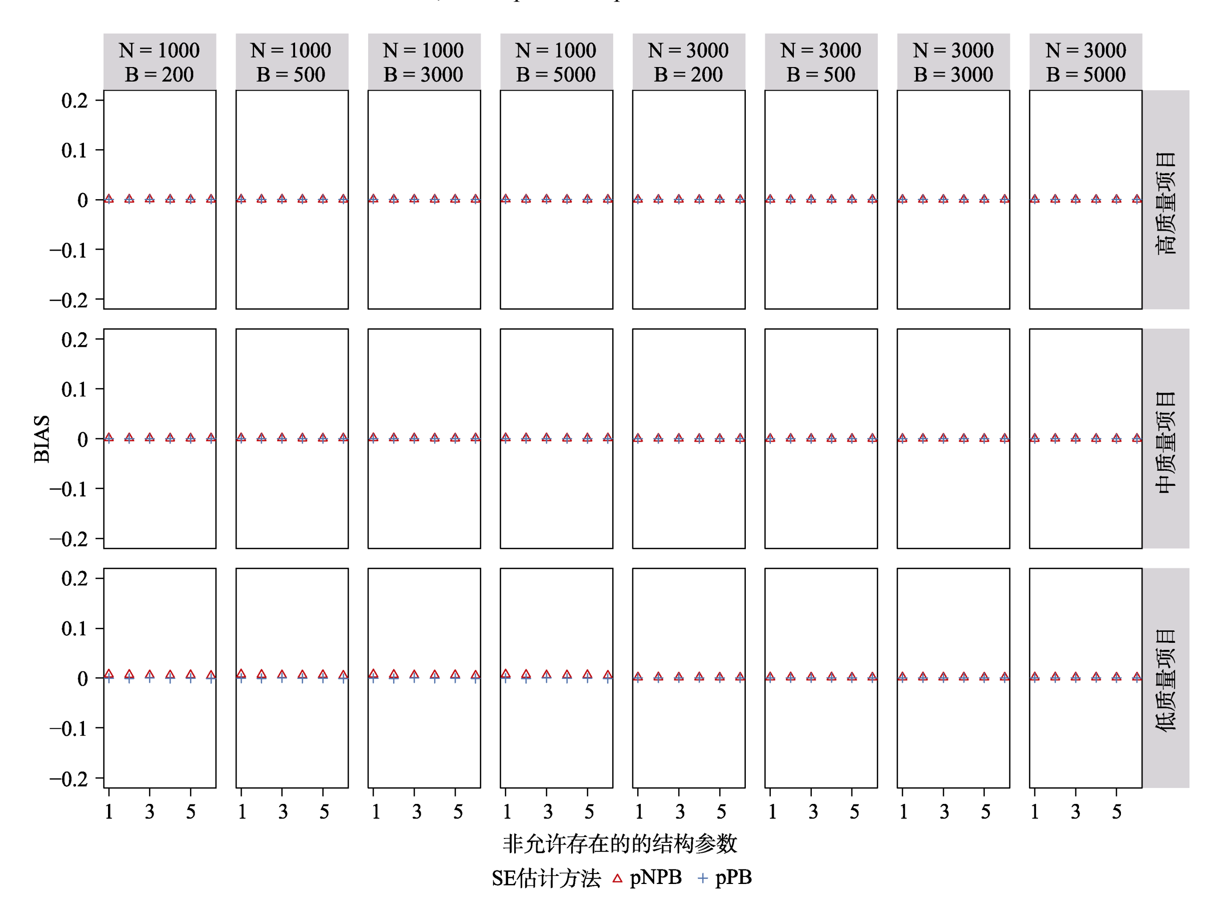


图 17　CDM 模型参数冗余时，基于 pNPB 与 pPB 的非允许存在结构参数的 *SE* 的 BIAS

英语语法测验项目上的作答。英语测验的内容专家与心理测量专家合作研究认为: 在这个数据集中共有 3 个属性: α_1(词法句法规则, morphosyntactic rules)、α_2(整合规则, cohesive rules)以及 α_3 (词汇规则, lexical rules), 图 18 中呈现了 ECPE 数据集的 **Q** 矩阵(Templin & Hoffman, 2013); 并且这 3 个属性之间可能存在线性层级结构关系: $\alpha_3 \to \alpha_2 \to \alpha_1$ (Liu et al., 2021; Templin & Bradshaw, 2014; Wang & Lu, 2021)。先前研究发现结构参数的 *SE* 在探索属性层级关系时有重要价值, 因此本文以 ECPE 数据的结构参数的 *SE* 估计为例, 对比以往相关研究结果(Liu et al., 2021), 展示本研究的理论与实践价值。

5.1　数据分析方法

使用同一链接下的饱和 G-DINA 模型估计模型参数, 使用 pPB 以及 pNPB 估计模型参数的 *SE*, 并与 PB 以及 NPB 比较运算时间。使用 *GDINA* 软件包估计模型参数, 基于 XPD、Obs 及 Sw 的模型参数的 *SE* 估计代码改编自 *dcminfo* 软件包(Liu & Xin, 2017), 其余功能自编 R 代码实现, 在云主机运行上允许全部程序。特别说明的是: (1)在 ECPE 数据的饱和结构模型中共有 $L = 2^3 = 8$ 种属性掌握模式, 因为结构参数之和等于 1, 因此将第 8 个结构参数约束为 $\eta_8 = 1 - \sum_{l=1}^{7} \eta_l$ 。(2)理论上讲, 重抽样次数越多, 获得准确 *SE* 估计结果的可能性就越大, 在本例中增加了 $B = 10000$ 时使用 pPB 以及 pNPB 估计 *SE* 结果; 由于 PB 以及 NPB 耗时会特别长, 因此没有考察这两个方法的运行时间。

5.2　研究结果

图 19 中呈现了饱和结构模型中 8 种属性掌握模式及其对应的结构参数估计值。表 1 中呈现的是使用不同方法计算的图 19 中呈现的结构参数估计值所对应的 *SE*。对比使用不同方法计算的结构参数的 *SE* 估计值可以发现, 整体上使用 pPB 方法估计的 *SE* 与使用 XPD 方法估计的 *SE* 在数值上非常接近; 使用 pNPB 方法估计的 *SE* 与使用 Sw 方法估计的 *SE* 在数值上比较接近。对比 pNPB 方法与 pPB 方法可以发现, pNPB 估计的 *SE* 的值比 pPB 方法估计的值要大, 这与模拟研究中 CDM 模型参数冗余时允许存在的结构参数的 *SE* 及非允许存在结构参数的 *SE* 的结果是一致的。

当 ECPE 数据中存在线性层级关系 $\alpha_3 \to \alpha_2 \to \alpha_1$ 时, 第 2、3、6 个结构参数(图 19 中灰色部分)应该近似等于 0 (Templin & Bradshaw, 2014), 然而对于特定的结构参数而言, 如 $\hat{\eta}_6$=0.014 是否近似等于 0, 需要统计检验。Liu 等人(2021)分别使用 XPD、Obs 以及 Sw 方法计算结构参数的 *SE*, 即 $SE(\hat{\eta})$, 然后使用公式(1)中的 *z* 统计量检验结构参数估计值 $\hat{\eta}$ 的显著性。他们研究发现, 除了 Obs 方法无法计算第 2 个参数的 *SE* 外, 使用基于 XPD、Obs 以及 Sw 方法的 *SE* 计算的 *z* 统计量, 在经过显著性水平校正后均一致地证实了存在线性层级关系的结论。在 *z* 统计量计算公式中, 结构参数估计值 $\hat{\eta}$ 在各个方法

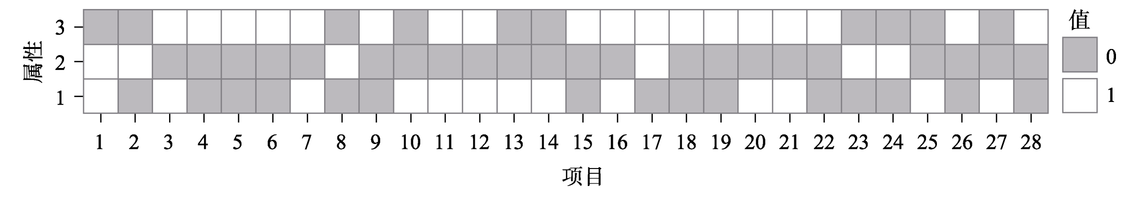


图 18　ECPE 数据集的 **Q** 矩阵

表 1　ECPE 数据的结构参数估计值的 *SE*

参数序号	解析法			pNPB				pPB			
	XPD	Obs	Sw	200	500	3000	10000	200	500	3000	10000
1	0.017	0.018	0.023	0.021	0.022	0.022	0.021	0.015	0.015	0.015	0.015
2	0.003	–	0.010	0.008	0.008	0.008	0.008	0.003	0.003	0.003	0.003
3	0.013	0.014	0.017	0.013	0.014	0.013	0.013	0.010	0.010	0.011	0.011
4	0.017	0.020	0.027	0.027	0.026	0.026	0.026	0.016	0.016	0.015	0.015
5	0.006	0.006	0.007	0.007	0.007	0.008	0.008	0.005	0.005	0.005	0.005
6	0.008	0.007	0.016	0.010	0.010	0.010	0.011	0.008	0.008	0.008	0.008
7	0.018	0.020	0.027	0.023	0.023	0.024	0.024	0.018	0.018	0.017	0.017

注: pNPB、pPB 对应的数字指重抽样次数。“–”表示无法计算。

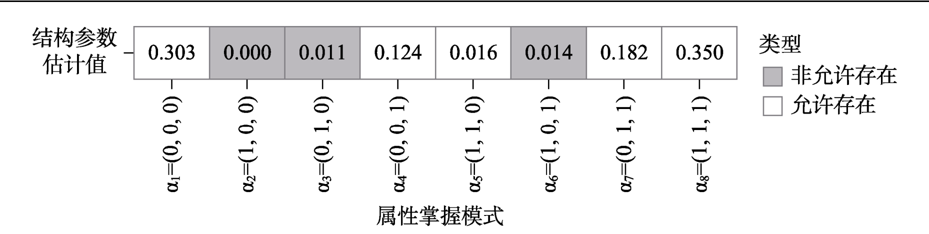


图 19　ECPE 数据集中所有可能的属性掌握模式及其对应的结构参数估计值

中是相同的，只有 $SE(\hat{\eta})$ 受到计算方法的影响而取值不同。本研究中，由于使用 pNPB 及 pPB 计算第 2、3、6 个结构参数的 *SE* 的值均处于使用 XPD、Obs 以及 Sw 方法计算的 *SE* 的最大值与最小值之间，所以使用公式(1)计算的 *z* 统计量的值也会位于解析法矩阵计算的 *z* 统计量的最大值与最小值之间。也就是，使用这两种方法计算的 *SE* 同样证实了线性层级关系的存在。需要明确指出的是，当 CDM 中存在属性层级关系时，XPD、Obs 以及 Sw 方法经常会遇到无法求逆的问题，而且对于 Obs 矩阵而言可能会由于计算误差的存在使得对角线元素小于 0 而无法计算 *SE* (如，本例中的第 2 个结构参数的 *SE*)。自助法使用通过重抽样数据估计获得的模型参数直接计算 *SE*，具有解析法所不具备的无需求逆矩阵的优点。另外，同模拟结果一致，在这个例子中同样可以发现增加重抽样次数对 *SE* 估计值产生了很小的影响，尤其是 $B \geq 3000$ 时。

为了直观地说明 pNPB 及 pPB 在运算效率上的提升，本文比较了使用 200、500 及 3000 次重抽样时新方法与传统自助法在计算时间上的差异。结果显示：pNPB 耗时分别是 10.93 s、25.43 s、135.36 s；pPB 耗时分别是 15.42 s、36.01 s、200.96 s；NPB 耗时分别是 158.43 s、392.97 s、2282.33 s；PB 耗时分别是 220.77 s、537.15 s、3201.17 s。可以发现，pNPB 及 pPB 极大地提升了计算效率。

6　讨论与展望

CDM 研究中，模型参数的 *SE* 及 CI 估计是一个具有重要价值且富有挑战性的问题(de la Torre, 2011; Liu et al., 2021; Ma & de la Torre, 2019; von Davier, 2014)。解析法信息矩阵 XPD、Obs 及 Sw 等在多数的应用情景中虽然有好的表现(Liu, Xin et al., 2019; Philipp et al., 2018; 刘彦楼 等, 2016)，但其缺点在于需要矩阵正定，且易受边界值问题的影响(DeCarlo, 2011, 2019)；传统自助法，如 NPB 以及 PB 虽然具有前提假设少、通用性强的优点，但是存在计算效率低、耗时长的问题(Ma & de la Torre, 2020b)。本研究提出使用 pNPB 以及 pPB 计算 CDM 模型参数的 *SE* 及 CI，系统探讨了模型设定、样本量、重抽样次数、项目质量及具体估计方法对 *SE* 及 CI 估计结果的影响；展示了 pNPB 以及 pPB 在分析可能存在属性层级关系的 CDM 实证数据 ECPE 时的检验效果与计算效率。

特别指出的是，除了解析法信息矩阵、自助法外还有其他方法可以用于计算 CDM 模型参数的 *SE* 与 CI，如 MCMC (Markov chain Monte Carlo)方法。MCMC 方法不仅可以用于计算模型参数估计值，而且可以通过计算估计过程中产生的模型参数的标准差，作为 *SE* 的估计。使用 MCMC 估计 CDM 的模型参数，计算耗时可能会特别长(例如，大于 1 小时)。对于模型参数的 *SE* 及 CI 进行研究时，需要进行大量的重复(如 500 次或以上)才能获得可靠的模拟结果(Liu, Xin et al., 2019; Philipp et al., 2018; 刘彦楼 等, 2016)。另外，这类基于贝叶斯的方法可能对于先验分布敏感(Jiang et al., 2021)。因此，本研究没有探讨使用 MCMC 算法计算 CDM 模型参数的 *SE* 及 CI 的表现。

6.1　讨论

(1)自助法在估计 *SE* 及 CI 时的表现

本质而言，无论是 NPB 还是 PB 都是模拟从总体中抽样获得样本数据的过程：将样本或通过样本估计获得的模型参数认为是“总体”再抽样计算的，是对于“样本”的再抽样。也就是，自助法无法超越它所依赖的“样本”而凭空产生出更多的信息。因此，在 CDM 的观察数据中所包含的关于未知参数的信息越多、越准确，自助法的效果会越好。模拟研究中发现，模型设定、样本量以及项目质量对于 pNPB 及 pPB 的表现有重要影响。这主要是因为在模型正确设定条件下，观察数据与模型是完美拟合的；而模型参数冗余条件下的情景与此相反，可以明显地观察到使用饱和模型拟合带有属性层级关系的数据时，由于非允许参数的存在，模型参数估计值的估计准确性受到了很大的影响。这从侧面说明了在 CDM 中进行属性层级关系检验或探索的重

要性(Hu & Templin, 2020; Liu et al., 2021; Ma & Xu, 2021)。样本量越大，所包含的关于未知参数的信息越多，模型参数估计值就会越准确；项目质量越高，越能有效区分被试的属性掌握模式状况，也就是说此时样本能够提供更多信息，从而使得 pNPB 及 pPB 的表现越好。通过模拟数据观察到的一个有意思的现象是在低质量项目条件下，与同实验水平组合的前半段参数相比，后半段的项目参数的 95% CI 覆盖率及 BIAS 的表现明显变差。观察 **Q** 矩阵可以发现，在最后 4 个项目中每个项目都测量了 3 个属性，也就是说每个项目中都有 8 个项目参数需要估计，也就是在低质量项目条件下最后的 4 个项目中可供利用的信息明显少于其他项目。

(2)重抽样次数对于自助法的影响

自助法是计算密集型方法，特定计算环境中重抽样次数越多计算时间也就会越长(Efron & Tibshirani, 1993)，就理论而言，重抽样次数的增加会增加 *SE* 估计准确的可能性(Hayes, 2009, 2018)。如前所述，在自助法中如何确定重抽样次数还没有明确的结论(Bai et al., 2016; Guo & Wind, 2021; Lai, 2021)。本研究在使用并行自助法计算效率提升的基础上，探索了 $B = 200$、500、3000 及 5000 时的表现。从整体而言，重抽样次数对于 pNPB 及 pPB 表现的影响较小，当重抽样次数 $B \geq 500$ 时各条件组合下的模拟结果开始变得稳定，$B = 3000$ 与 $B = 5000$ 两种重抽样次数下的结果则几乎完全相同。模型完全正确设定时一些条件下的参数或模型冗余设定时允许存在参数的 95% CI 覆盖率及 BIAS 的表现随着重抽样次数 *B* 从 200 增加到 3000 稍有变好；在一些非理想情景下，如项目质量低、非允许存在参数等，重抽样次数的增加对于 pNPB 及 pPB 表现没有明显影响。实证数据分析发现 pNPB 在 200、500 和 3000 下的结果与 10000 次重抽样次数下的结果相比仅有细微的差别，pPB 在 3000 次重复时的结果与 10000 次重复下的结果几乎一致。理论上而言，CDM 的信息矩阵是关于观察数据中包含的模型参数信息的度量(Liu, Xin et al., 2019)，而 *SE* 则是关于模型参数估计值不确定信息的度量(Liu et al., 2021)，这也就是说，观察数据中包含“信息”量的多少是影响 *SE* 表现的主要因素。本文的模拟及实证研究支持以上理论，因此作者认为影响自助法表现的最主要因素并非重抽样次数，而是观察数据中所包含“信息”的多少。当然，本文结论是否可以推广到其他情景中有待进一步研究。

6.2　研究展望

有一些重要问题需要在后续研究中进一步探讨。(1)本文仅在项目数量为 30，属性数量为 4 的条件下展开研究，后续研究者可以继续探讨不同项目数量及属性数量对于 pNPB 及 pPB 的影响。(2)本研究仅以($\alpha_1 \rightarrow \alpha_2$，$\alpha_1 \rightarrow \alpha_3$)层级关系为例，探讨了模型参数冗余设定对于 pNPB 及 pPB 表现的影响，然而不同属性层级关系条件下，模型参数的 *SE* 的表现，尤其是结构参数的 *SE* 的表现有待进一步探索。现实中不仅会存在属性层级关系，而且可能会同时存在属性之间的相关(Hu & Templin, 2020; Liu et al., 2021)，限于研究目的，本研究没有考虑这种情景。本文认为 pNPB 及 pPB 在探索及验证属性层级关系时的表现值得进一步研究。(3)除了本研究中使用的模型参数 95% CI 计算方法外，还有一些基于自助法的 CI 计算方法的表现也值得进一步关注(例如，Jiang, 2021; Lai, 2021)。(4)解析法信息矩阵在属性层级关系存在时经常会遇到无法求逆的问题，因此本研究无法直接比较这两类方法的优劣，Liu 等人(2021)初步提出了通过逐步排除非允许存在结构参数的两阶段模型参数估计的思路，这也是一个具有重要理论及实践价值的方向。本研究在 CDM 模型参数完全正确设定条件下对比了解析法 XPD、Obs、Sw、pNPB 及 pPB 的表现，结果显示，解析法(如，Obs 或 Sw)在一些条件下的表现要稍优于 pNPB 或 pPB。后续研究可以比较两阶段模型参数估计思路下的解析法与 pNPB 及 pPB 方法的表现。(5)需要特别指出的是，pNPB 及 pPB 除可以用于计算 *SE* 及 CI 外，还有很多潜在的理论及实践价值。研究者可以进一步探索 pNPB 及 pPB 在项目功能差异检验、项目水平上的模型比较、**Q** 矩阵检验等领域中的表现。(6)本文在 CDM 框架下探讨了 pNPB 及 pPB 的表现，但是作为通用性强的一类方法，后续研究者可以在开发并行方法的基础上，在其他统计与测量模型中深入探讨自助法的表现，以解决先前研究没有明确的结论或结论相冲突的问题(例如，Efron & Tibshirani, 1993; Hayes, 2009, 2018; Lai, 2021)。

7　结论

结果显示：(1) CDM 完全正确设定时，在高质量及中等质量项目条件下，使用 pNPB 及 pPB 这两种方法计算的项目参数和结构参数 95% CI 覆盖率及 BIAS 均有好的表现；且随着样本量的增大及项

目质量的变好，这两种方法的表现也在变好。低项目质量严重影响了 pNPB 及 pPB 的表现，pNPB 倾向于高估模型参数的 *SE*, pPB 则倾向于低估 *SE*。(2)在 CDM 的模型参数存在冗余时，在高质量及中等质量项目条件下，使用 pNPB 及 pPB 这两种方法计算的大部分允许存在项目参数和几乎全部允许存在结构参数的 95% CI 覆盖率及 BIAS 均有好的表现，但是也存在部分项目参数的 95% CI 覆盖率极端偏离理论区间且 BIAS 值为负数的情况。非允许存在项目参数及结构参数的 95% CI 覆盖率在大多数条件下表现较差。(3)探讨了 pNPB 及 pPB 在实证数据中的效果，发现使用 pNPB 及 pPB 计算的 *SE*, 获得了同先前研究一致的结论，即 ECPE 数据中存在线性属性层级关系；同 NPB 及 PB 相比, pNPB 及 pPB 极大地提升了计算效率，是简易、可行的 *SE* 及 CI 计算方法。(4)综合模拟研究与实证数据分析结果，本研究初步认为：在 pNPB 及 pPB 方法中为快速预览 *SE* 估计结果可以选择 200 次重抽样；为获得较为准确的估计结果，审慎起见可以选择 3000 或以上的重抽样次数。

参 考 文 献

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). Washington.

Bai, H., Sivo, S. A., Pan, W., & Fan, X. (2016). Application of a new resampling method to SEM: A comparison of S-SMART with the bootstrap. *International Journal of Research & Method in Education, 39*(2), 194–207.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (2007). *Discrete multivariate analysis: Theory and practice*. Springer.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge.

DeCarlo, T. (2019). Insights from reparameterized DINA and beyond. In M. von Davier & Y.-S. Lee (Eds.). *Handbook of diagnostic classification models* (pp. 549–572). Springer.

DeCarlo, T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-Matrix. *Applied Psychological Measurement, 35*(1), 8–26.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

de la Torre, J., & Lee, Y. S. (2013). Evaluating the wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*(4), 355–373.

Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software, 71*(9), 1–25.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Guo, W., & Wind, S. A. (2021). An iterative parametric bootstrap approach to evaluating rater fit. *Applied Psychological Measurement, 45*(5), 315–330.

Gu, Y., & Xu, G. (2019). Learning attribute patterns in high-dimensional structured latent attribute models. *Journal of Machine Learning Research, 20*(115), 1–58.

Gu, Y., & Xu, G. (2020). Partial identifiability of restricted latent class models. *The Annals of Statistics, 48*(4), 2082–2107.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs, 76*(4), 408–420.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-Based approach (2nd ed.)*. Guilford.

Hu, B., & Templin, J. (2020). Using diagnostic classification models to validate attribute hierarchies and evaluate model fit in Bayesian networks. *Multivariate Behavioral Research, 55*(2), 300–311.

Jiang, Z., Raymond, M., DiStefano, C., Shi, D., Liu, R., & Sun, J. (2021). A Monte Carlo study of confidence interval methods for generalizability coefficient. *Educational and Psychological Measurement*. Advance online publication. https://doi.org/10.1177/00131644211033899

Khorramdel, L., Shin, H. J., & von Davier, M. (2019). GDM Software *mdltm* Including Parallel EM Algorithm. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 603–628). Springer.

Lai, M. H. C. (2021). Bootstrap confidence intervals for multilevel standardized effect size. *Multivariate Behavioral Research, 56*(4), 558–578.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*(3), 205–237.

Liu, R. (2018). Misspecification of attribute structure in diagnostic measurement. *Educational and Psychological Measurement, 78*(4), 605–634.

Liu, Y., Andersson, B., Xin, T., Zhang, H., & Wang, L. (2019). Improved wald statistics for item-level model comparison in diagnostic classification models. *Applied Psychological Measurement, 43*(5), 402–414.

Liu, Y., & Maydeu-Olivares, A. (2014). Identifying the source of misfit in item response theory models. *Multivariate Behavioral Research, 49*(4), 354–371.

Liu, Y., Tian, W., & Xin, T. (2016). An application of M_2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics, 41*(1), 3–26.

Liu, Y., & Xin, T. (2017). *dcminfo: Information matrix for diagnostic classification models*. R package version 0.1.6. https://CRAN.R-project.org/package=dcminfo

Liu, Y., Xin, T., Andersson, B., & Tian, W. (2019). Information matrix estimation procedures for cognitive diagnostic models. *British Journal of Mathematical and Statistical Psychology, 72*(1), 18–37.

Liu, Y., Xin, T., & Jiang, Y. (2021). Structural parameter standard error estimation method in diagnostic classification models: Estimation and application. *Multivariate Behavioral Research.* Advance online publication. https://doi.org/10.1080/00273171.2021.1919048

Liu, Y., Xin, T., Li, L., Tian, W., & Liu, X. (2016). An improved method for differential item functioning

detection in cognitive diagnosis models: An application of wald statistic based on observed information matrix. *Acta Psychologica Sinica, 48*(5), 588–598.
[刘彦楼, 辛涛, 李令青, 田伟, 刘笑笑. (2016). 改进的认知诊断模型项目功能差异检验方法——基于观察信息矩阵的 Wald 统计量. *心理学报, 48*(5), 588–598.]
Liu, Y., Yin, H., Xin, T., Shao, L., & Yuan, L. (2019). A comparison of differential item functioning detection methods in cognitive diagnostic models. *Frontiers in Psychology*, *10*, 1137.
Ma, C., & Xu, G. (2021). Hypothesis testing for hierarchical structures in cognitive diagnosis models. *arXiv preprint arXiv:2106.03218v1*
Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69*(3), 253–275.
Ma, W., & de la Torre, J. (2019). Category-level model selection for the sequential G-DINA model. *Journal of Educational and Behavioral Statistics, 44*(1), 45–77.
Ma, W., & de la Torre, J. (2020a). An empirical Q‐matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology, 73*(1), 142–163.
Ma, W., & de la Torre, J. (2020b). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software, 93*(14), 1–26.
Ma, W., Ragip, T., & de la Torre, J. (2021). Detecting differential item functioning using multiple-group cognitive diagnosis models. *Applied Psychological Measurement, 45*(1), 37–53.
Philipp, M., Strobl, C., de la Torre, J., & Zeileis, A. (2018). On the estimation of standard errors in cognitive diagnosis models. *Journal of Educational and Behavioral Statistics, 43*(1), 88–115.
Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2020). *CDM: cognitive diagnosis modeling*. R package version 7.5-15. http://CRAN.R-project.org/package=CDM
Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: theory, methods, and applications*. Guilford.
Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*(2), 317–339.
Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice, 32*(2), 37–50.
Tjoe, H., & de la Torre, J. (2014). On recognizing proportionality: Does the ability to solve missing value proportional problems presuppose the conception of proportional reasoning? *The Journal of Mathematical Behavior, 33*, 1–7.
von Davier, M. (2014). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 49–71.
Wang, C., & Lu, J. (2021). Learning attribute hierarchies from data: Two exploratory approaches. *Journal of Educational and Behavioral Statistics, 46*(1), 58–84.
Wu, Z., Deloria-Knoll, M., & Zeger, S. L. (2017). Nested partially latent class models for dependent binary data; estimating disease etiology. *Biostatistics, 18*(2), 200–213.
Zhang, Z. (2014). Monte Carlo based statistical power analysis for mediation models: methods and software. *Behavior research methods, 46*(4), 1184–1198.
Zhang, Z., & Wang, L. (2020). *bmem: mediation analysis with missing data using Bootstrap*. R package version 1.8. https://CRAN.R-project.org/package=bmem

Standard errors and confidence intervals for cognitive diagnostic models: Parallel bootstrap methods

LIU Yanlou

(*Academy of Big Data for Education, Qufu Normal University, Jining* 273165, *China*)

Abstract

The model parameter standard error (*SE*; or variance-covariance matrix), which provides an estimate of the uncertainty associated with the model parameter estimate, has both theoretical and practical implications in cognitive diagnostic models (CDMs). The drawbacks of the analytic methods, such as the empirical cross-product information matrix, observed information matrix, and “robust” sandwich-type information matrix, are that they require the positive definiteness of the information matrix and may suffer from boundary problems. Another method for estimating model parameter *SE*s is to use the computer-intensive bootstrap method, and consequently, no study has systematically explored the performance of the bootstrap in calculating model parameter *SE*s and confidence intervals (CIs) in CDMs.

The purpose of this research is to present two new highly efficient bootstrap methods to calculate model parameter *SE*s and CIs in CDMs, namely the parallel parametric bootstrap (pPB) and parallel non-parametric bootstrap (pNPB) methods. A simulation study was conducted to evaluate the performance of the pPB and pNPB methods. Five factors that may influence the performance of the model parameter *SE*s and CIs were manipulated. The two model specification scenarios considered in this simulation were the correctly specified and

over-specified models. The sample size was set to two levels: 1, 000 and 3, 000. Three bootstrap sample sizes were manipulated: 200, 500, and 3, 000. Three levels of item quality were considered: high [$P(\mathbf{0})=0.1$, $P(\mathbf{1})=0.9$], moderate [$P(\mathbf{0})=0.2$, $P(\mathbf{1})=0.8$], and low quality [$P(\mathbf{0})=0.3$, $P(\mathbf{1})=0.7$]. The pPB and pNPB methods were used to estimate model parameter *SE*s and CIs.

The simulation results indicated the following.

(1) For the correctly specified CDMs, under the high- or moderate-item-quality conditions, the coverage rates of the 95% CIs of the model parameter *SE*s based on the pNPB or pPB method were reasonably close to the expected coverage rate, and the bias for each model parameter *SE* converged to zero, meaning that the estimated *SE* was almost identical to the empirical *SE*. The increase in the bootstrap sample size had only a slight effect on the performance of the pNPB or pPB method. Under the low-item-quality condition, the pNPB method tended to over-estimate *SE*, whereas a contrary trend was observed for the pPB method.

(2) For the over-specified CDMs, most of the permissible item parameter *SE*s and almost all of the permissible structural parameter *SE*s exhibited good performance in terms of the 95% CI coverage rates and bias. Under most of the simulation conditions, the impermissible model parameter *SE*s did not exhibit good performance in approximating the empirical *SE*s.

To the best of our knowledge, this is the first study in which the performance of the bootstrap method in estimating model parameter *SE*s and CIs in CDMs is systematically investigated. The pNPB or pPB appears to be a useful tool for researchers interested in evaluating the uncertainty of the model parameter point estimates. As a time-saving computational strategy, the pNPB or pPB method is substantially faster than the usual bootstrap method. The simulation and real data studies showed that 3, 000 re-samples might be adequate for the bootstrap method in calculating model parameter *SE*s and CIs in CDMs.

Key words cognitive diagnostic model, standard error, confidence interval, bootstrap, parallel computing method